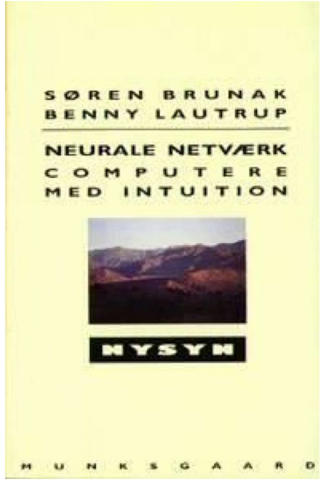
Identifikation af diskriminative features i sundhedsdata ved brug af machine learning

Søren Brunak

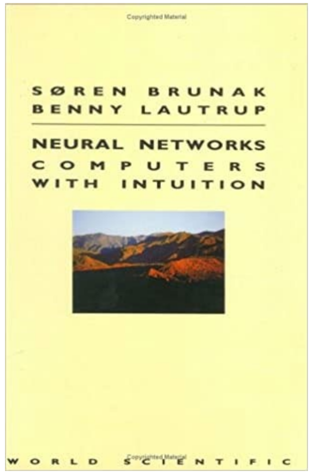
Novo Nordisk Foundation Center for Protein Research
University of Copenhagen
soren.brunak@cpr.ku.dk

Rigshospitalet
soeren.brunak@regionh.dk

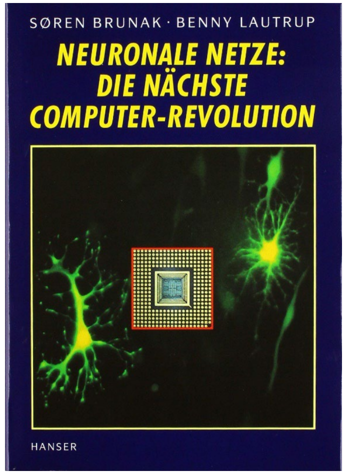
Machine learning bøger



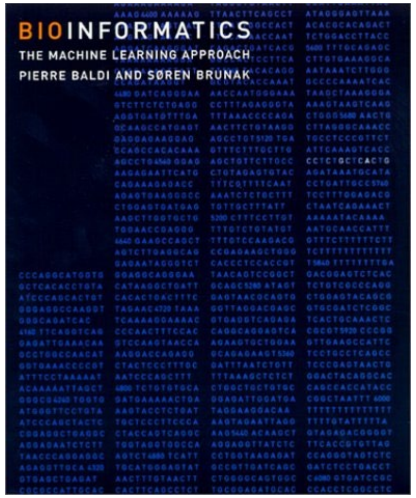
Dansk 1988



Engelsk 1990

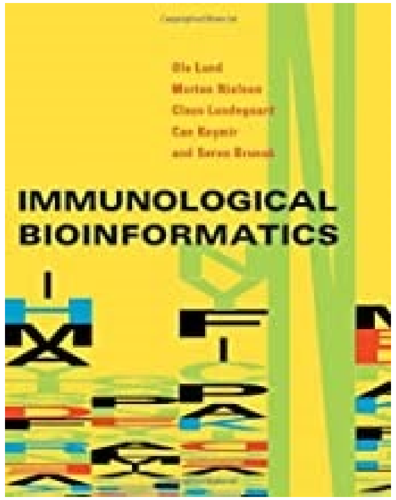


Tysk 1993



1998

2001



2005

MIT Press

AP photo

Elementary school teachers picket against use of calculators in grade school

The teachers feel if students use calculators too early, they won't learn math concepts

Math teachers protest against calculator use

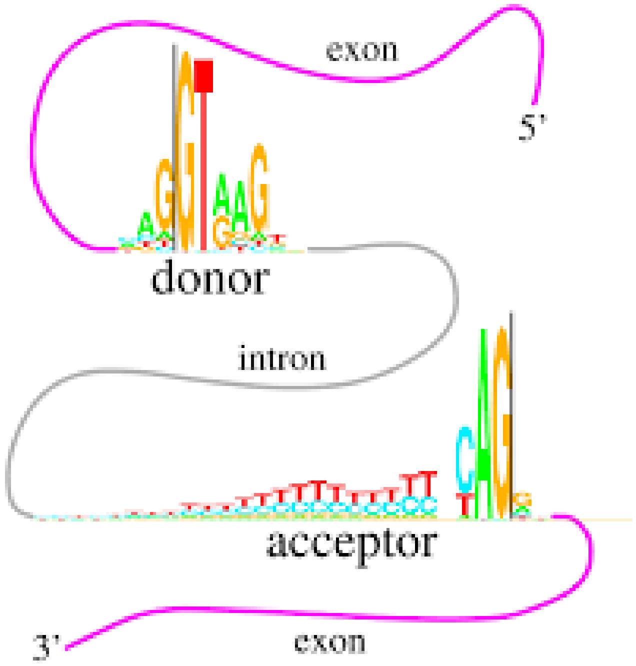
By JILL LAWRENCE

"My older kids don't pay any attention to an answer being absurd.

strate," he said. "Teachers are shy."

Data versus Method

Data versus Method



Cleaning up gene databases

SIR—We have discovered errors in the EMBL nucleotide sequence databank in the course of training of artificial neural networks to recognize pre-messenger RNA splicing signals[1] in human genes. In training on thirty-three human genes, the seven genes listed in the table appeared to disturb the learning process. Further investigation revealed discrepancies between the databank sequences and the original published papers for three of the genes. In the other four examples, we found wrongly assigned splicing frames of introns.

The subset of thirty-three human genes from the EMBL databank carried information about the location of splicing signals. On the basis of the feature table information, we assigned each nucleotide to one of two categories, splicing donor sites or otherwise. We used these data to train a neural network of the perceptron type[2] to classify nucleotides, on the basis of their context, as belonging to one of the two categories. The network was similar to that used for predicting secondary structures of proteins[3,4].

Neural networks are now being applied to many classification tasks; they have in common the ability, when trained, of dealing with non-linearities in the association between objects and categories. But in the presence of strong non-linearities, as when many similar objects must be put into widely different categories while many dissimilar objects are destined for the same category, the classification can usually be learned by the network only with difficulty. Errors introduced randomly lower the degree of regularity, which is also a prerequisite for the function of a specific biological recognition mechanism.

The problem of errors in the sequences incorporated in databanks could be amplified in the future, as a consequence of the increasing rate at which sequences are determined and by the initiative to sequence the human genome. We therefore propose that computerized proof reading should be incorporated in the databanks, so as to take advantage of the ability of neural networks to reveal non-linearities in a dataset, as we have demonstrated.

Errors in the EMBL nucleotide sequence database

Databank sequence (Accession no.)	Published sequence	Type of error
HSHLIA (X00492)	ref. 5	Splice donor cited at base pair (bp) 328 in feature table (FT). Correct position: 329 (there is a typographical ambiguity in the article)
		Splice donor cited at bp 2971 in FT. Correct position 2981
HSBGL3(V00499)	ref. 6	Presumed −1 bp error in splicing frame of intron I in article*
HSERPG (X02158)	ref. 7	Nucleotides CG missing after bp 578
		All following FT entries −2 bp wrong until last acceptor site which is +3 bp wrong
HSDGL1(V00505)	ref. 8	Presumed +1 bp error in splicing frame of intron I in article*
HSEGL1(V00510)	ref. 9	Presumed +1 bp error in splicing frame of intron I in article*
HSHSP27(X03900)	ref. 10	Presumed −1 bp error in splicing frame of intron I in article*
HSGROW2(V00520)	ref. 11	Splice donor cited at position 1081 in FT. Correct position: 1091

Discrepancies between published and databank entered sequences, or presumed errors in the published sequences.

* The occurrence of repeated bases around splice donors and splice acceptors means that the splicing frame of an intron cannot always be unambiguously assigned on the basis of the messenger RNA sequence and amino-acid sequence. A neural network can, however, point out the frame by recognition of the splice signals present in the sequence. Splicing always occurs at a position that is fixed with respect to the sequence signal[12].

SØREN BRUNAK
Department of Structural Properties of Materials,
The Technical University of Denmark,
DK-2800 Lyngby,
Denmark

JACOB ENGELBRECHT
Department of Dairy Science,
Royal Veterinary and Agricultural University,
Bülowsvej 13,
DK-1870, Frederiksberg C,
Denmark

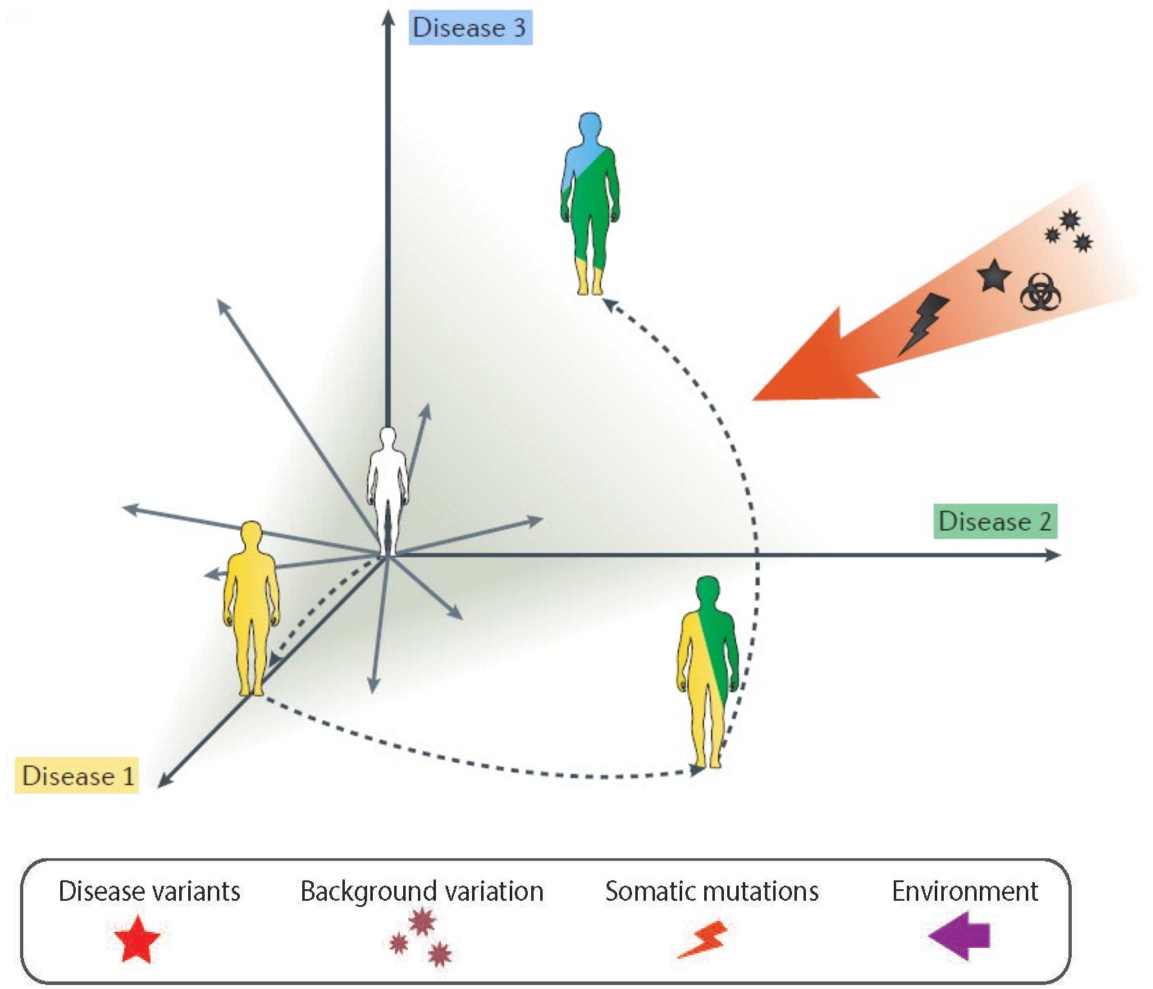
STEEN KNUDSEN
University Institute of Microbiology,
Øster Farimagsgade 2A,
DK-1353 Copenhagen C,
Denmark

1. Padgett, R.A., Grabowski, P.J., Konarska, M.M., Sheiler, S. & Sharp, P.A. *A. Rev. Biochem.* **55**, 1119–1150 (1986).
2. Rumelhart, D.E., Hinton, G.E. & Williams, R.J. in *Parallel Distributed Processing* (eds McClelland, J.L., Rumelhart, D.E. and the PDP Research Group) 318–362 (MIT, Cambridge, Massachusetts, 1987).
3. Qian, N. & Sejnowski, T.J. *J. molec. Biol.* **202**, 865–884 (1988).
4. Bohr, H. *et al. FEBS Lett.* **241**, 223–228 (1988).
5. Strachan, T., Sodoyer, R., Damotte, M. & Jordan, B.R. *EMBO J* **3**, 887–894 (1984).
6. Lawn, R.M., Efstratiadis, A., O'Connell, C. & Maniatis, T. *Cell* **21**, 647–651 (1980).
7. Jacobs, K. *et al. Nature* **313**, 806–810 (1985).
8. Spritz, R.A., DeRiel, J.K., Forget, B.G. & Weissman, S.M. *Cell* **21**, 639–646 (1980).
9. Baralle, F.E., Shoulders, C.C. & Proudfoot, N.J. *Cell* **21**, 621–626 (1980).
10. Hickey, E. *et al. Nucleic Acids Res.* **14**, 4127–4145 (1986).
11. DeNoto, F.M., Moore, D.D. & Goodman, H.M. *Nucleic Acids Res.* **9**, 3719–3730 (1981).
12. Jacob, M. & Gallinaro, H. *Nucleic Acids Res.* **17**, 2159–2175 (1989).

Brunak et al.
Nature **343**:123 (1990)

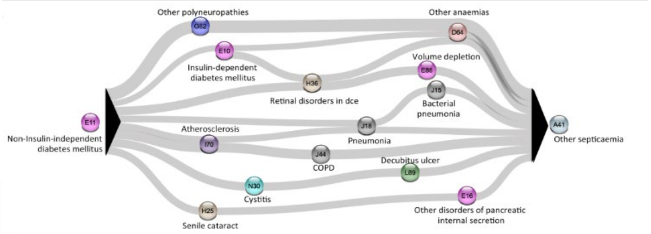
HUNKAT

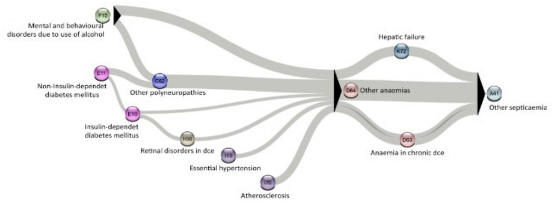
Lifelong multimorbidity journeys in disease space

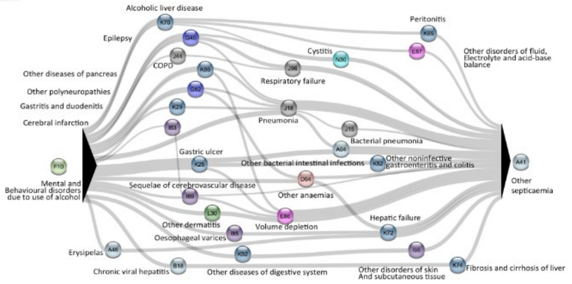


Hu, Thomas & Brunak
Nature Rev. Genetics 2016

The route towards disease impacts risks and outcomes





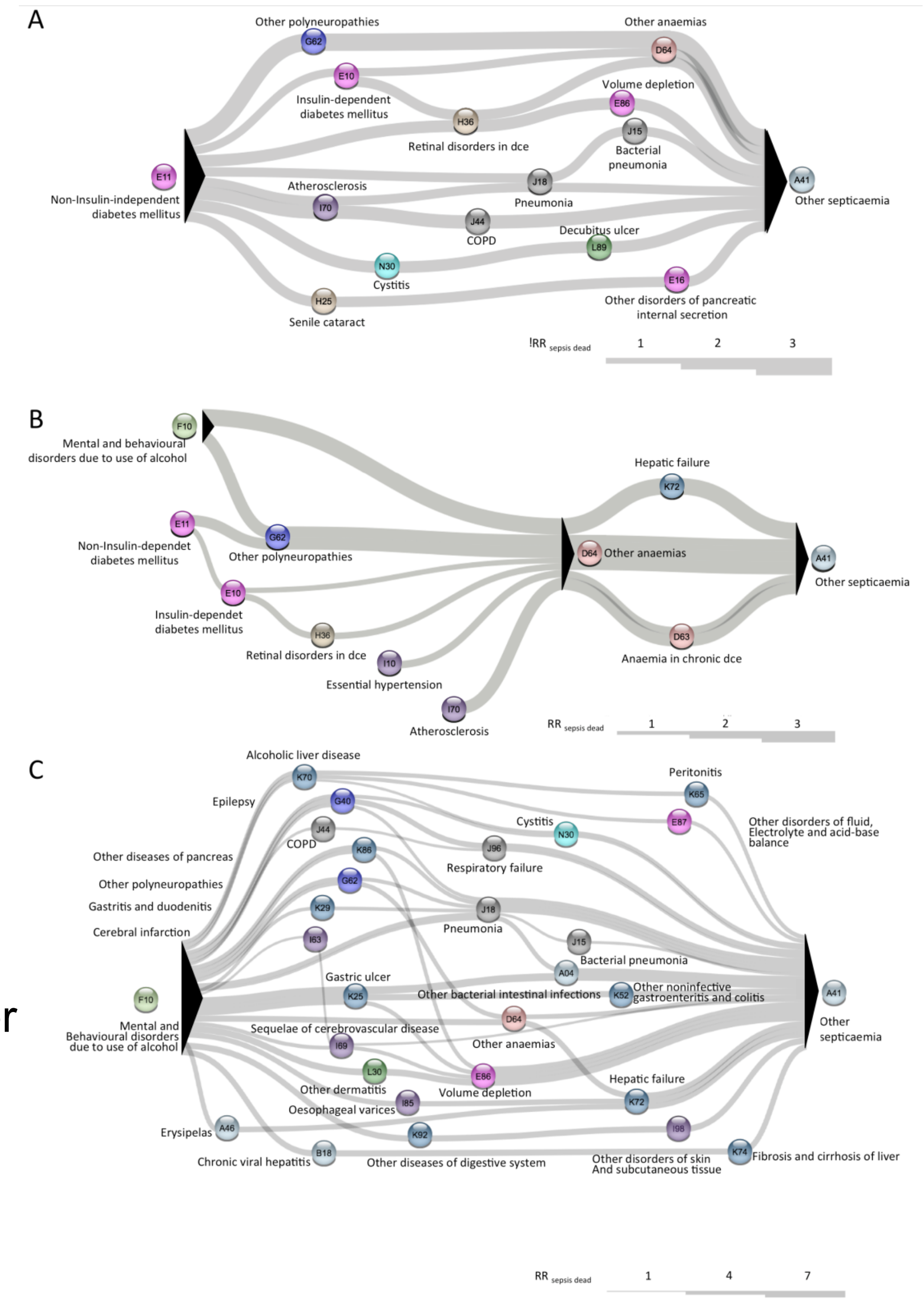


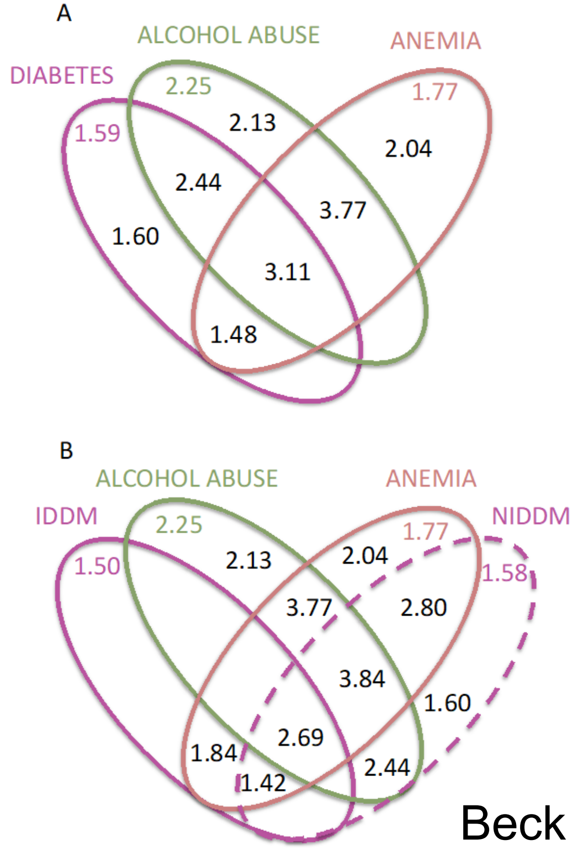
Diagnosis trajectories of prior multi-morbidity predict sepsis mortality, Beck et al. Sci .Rep. 2016

Diabetes

Anaemia/
Cancer

Mental disorder





Beck et al, 2016

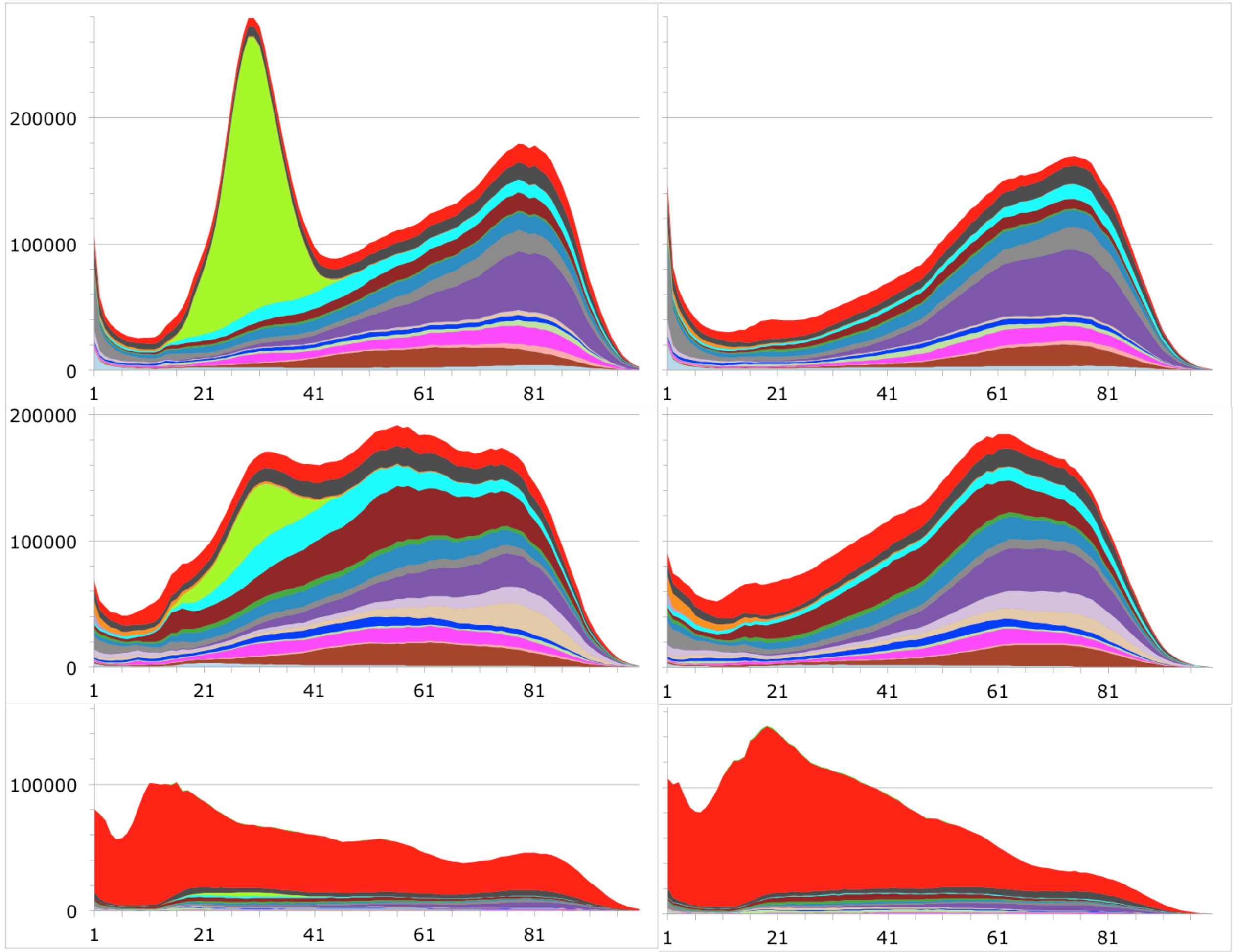
National Patient Registry (~7M Danes) ICD-10 diagnoses as a function of age

(ICD-10 era, 1994-2019)

In patients

Out patients

IR patients



Females

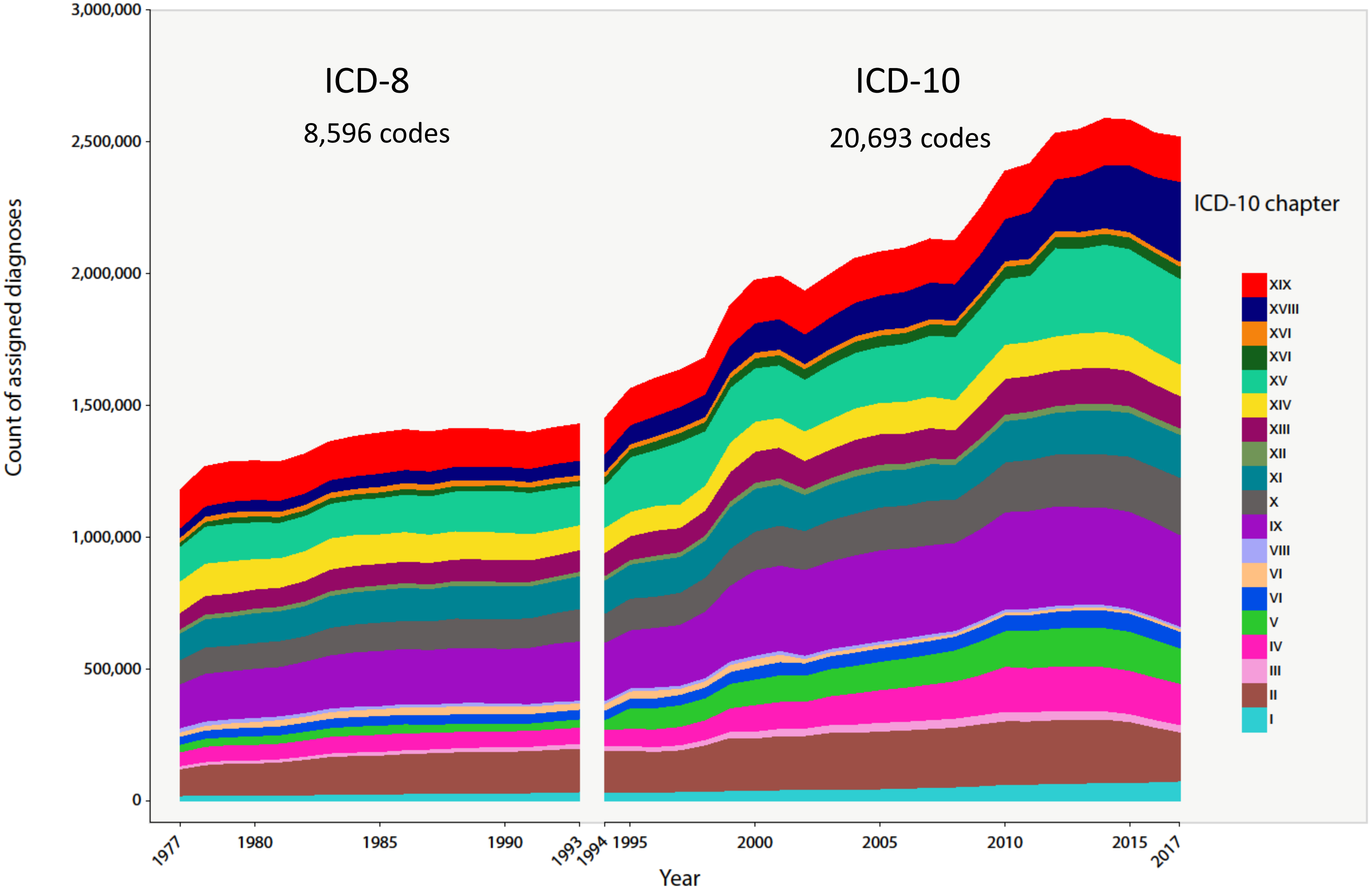
Males

ICD 10 chapter coloring
1: Certain infectious and parasitic diseases
2: Neoplasms
3: Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
4: Endocrine, nutritional and metabolic diseases
5: Mental and behavioural disorders
6: Diseases of the nervous system
7: Diseases of the eye and adnexa
8: Diseases of the ear and mastoid process
9: Diseases of the circulatory system
10: Diseases of the respiratory system
11: Diseases of the digestive system
12: Diseases of the skin and subcutaneous tissue
13: Diseases of the musculoskeletal system and connective tissue
14: Diseases of the genitourinary system
15: Pregnancy, childbirth and the puerperium
16: Certain conditions originating in the perinatal period
17: Congenital malformations, deformations and chromosomal abnormalities
18: Symptoms, signs and abnormal clinical and aboratory findings, not elsewhere classified
19: Injury, poisoning and certain other consequences of external causes
20: External causes of morbidity and mortality

AB Jensen et al., Nature Comm., 2014

Danish population-wide diagnoses from 1977-2019

ICD-8/ICD-10 periods, 9.5 million patients in the national registry

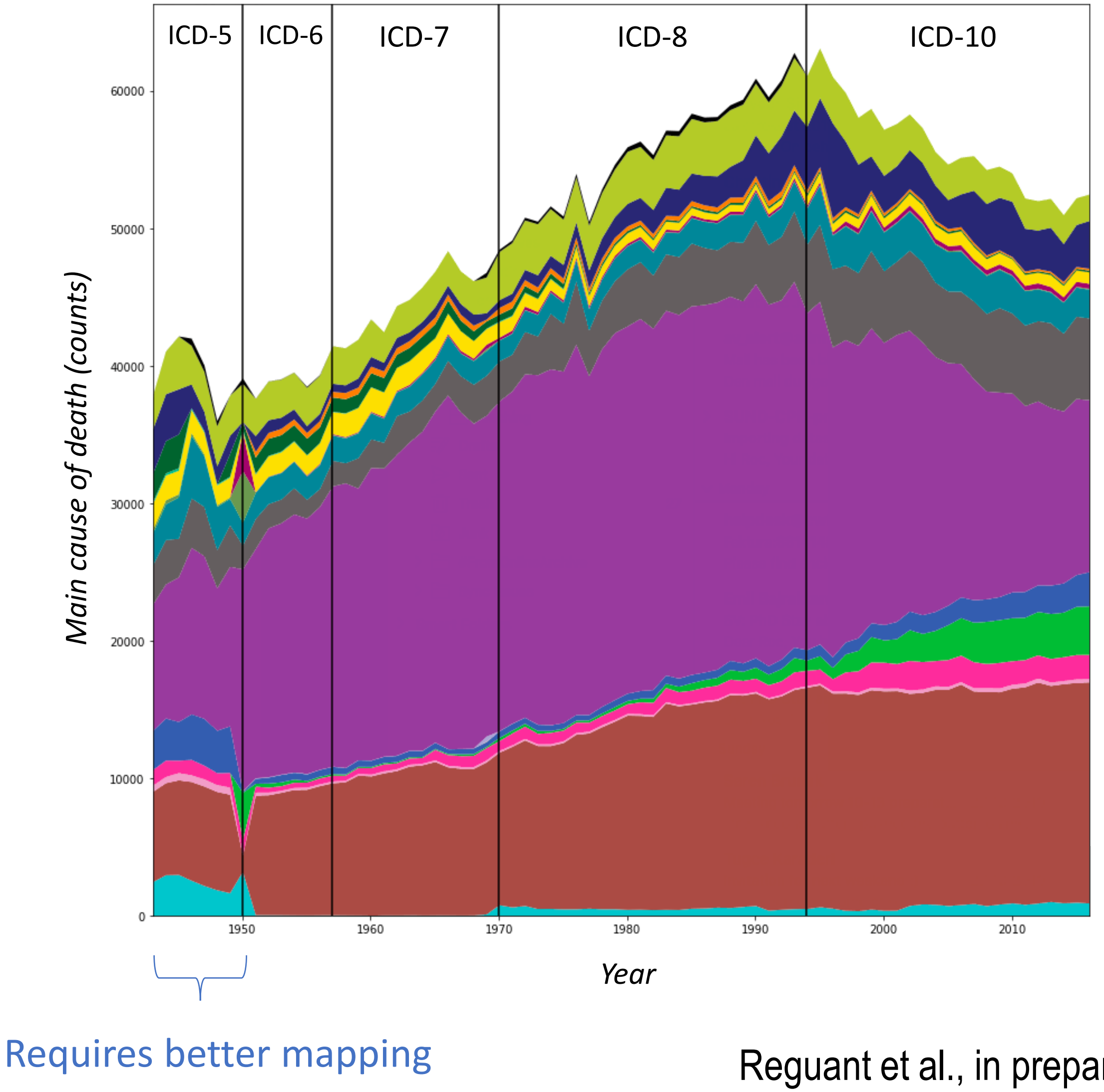


Pedersen et el.
Eur. J. Epi. 2023,

Death registry data, 1943-2018

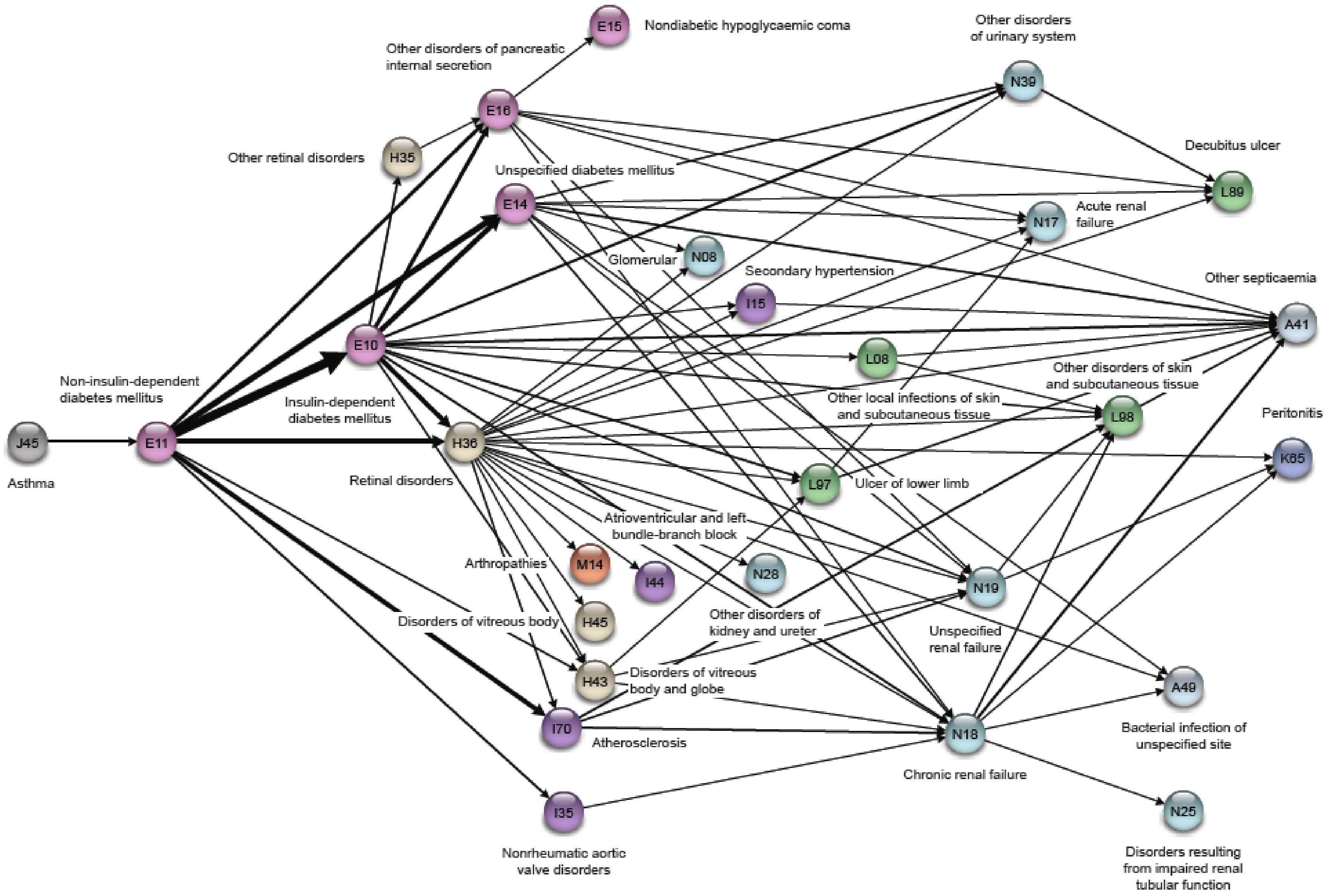
Death certificates from Denmark, ~75 years of data 1943-2018
Up to 8 contributing causes of death

Principal coding systems:
< 1994: ICD-8 (mapped to ICD-10)
>=1994: ICD-10



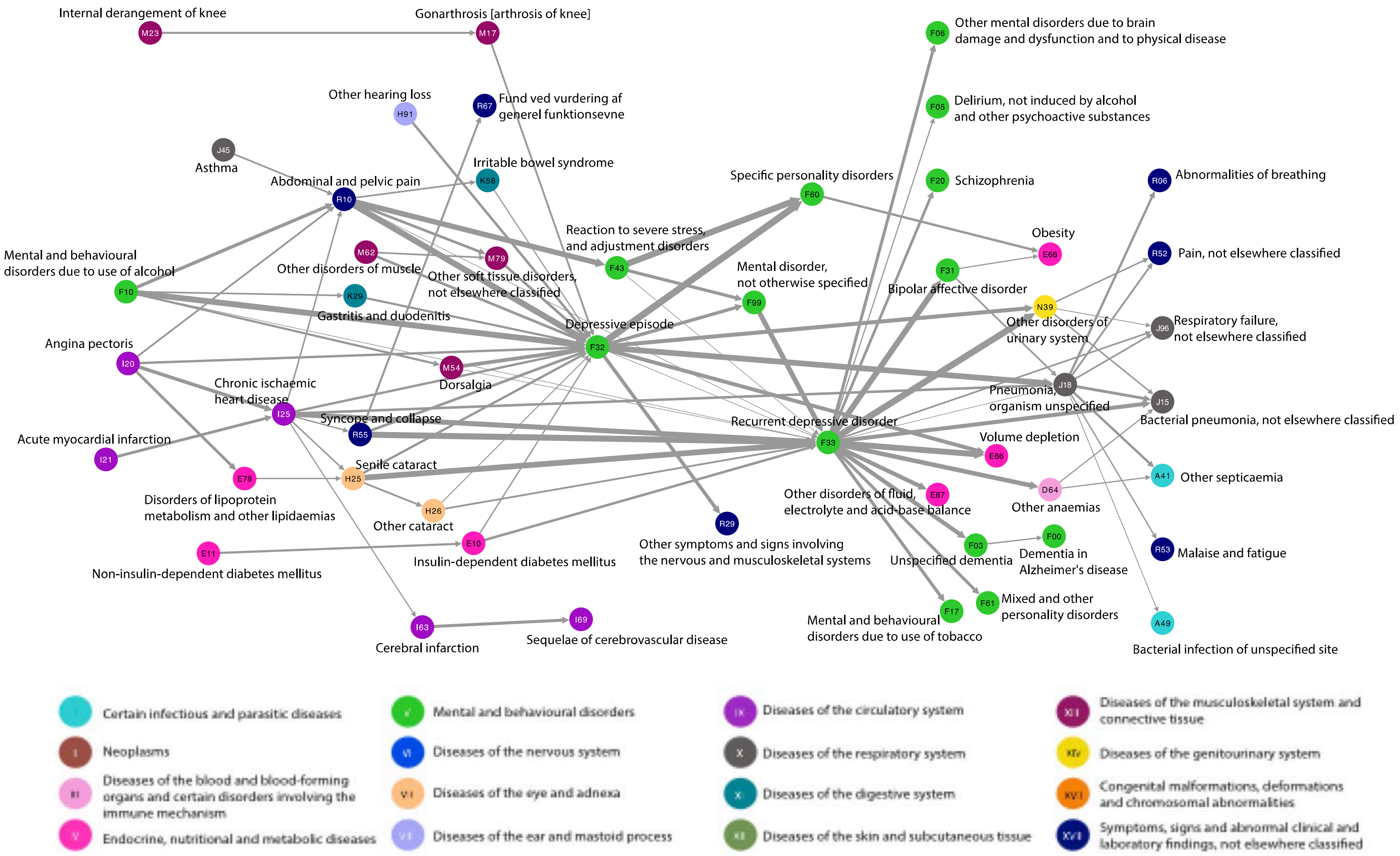
Reguant et al., in preparation 2023

Diabetes trajectory network

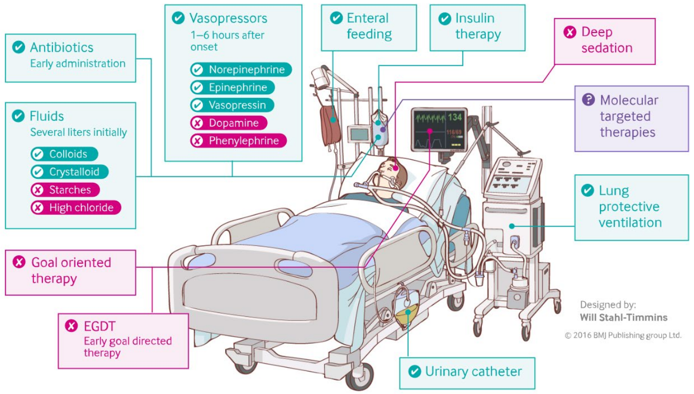


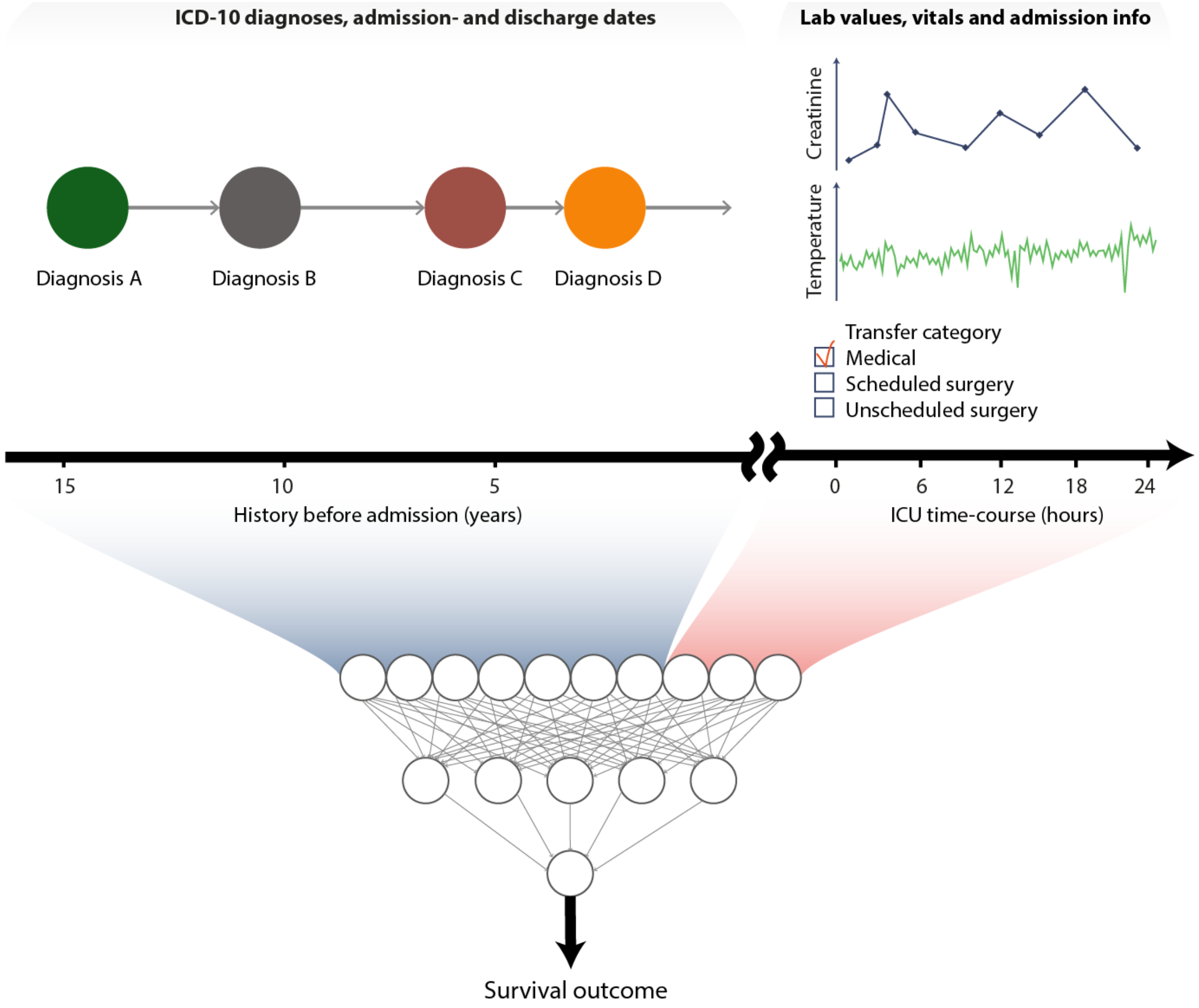
AB Jensen et al.,
Nature Comm., 2014

Disease trajectory network of depression patients



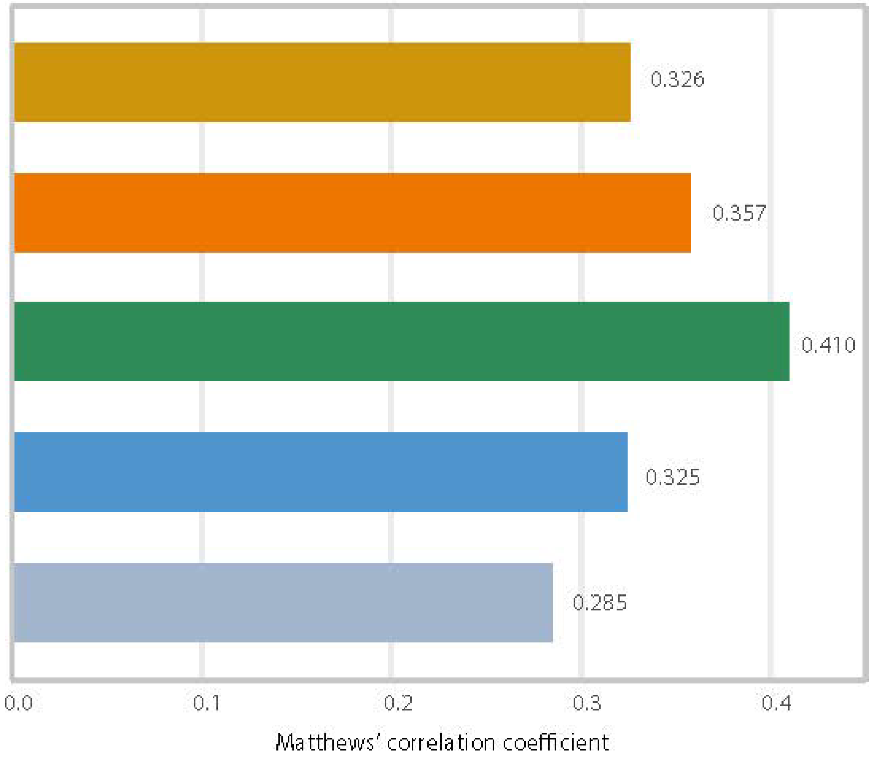
ICU patient mortality prediction from machine learning based aggregation of time scales

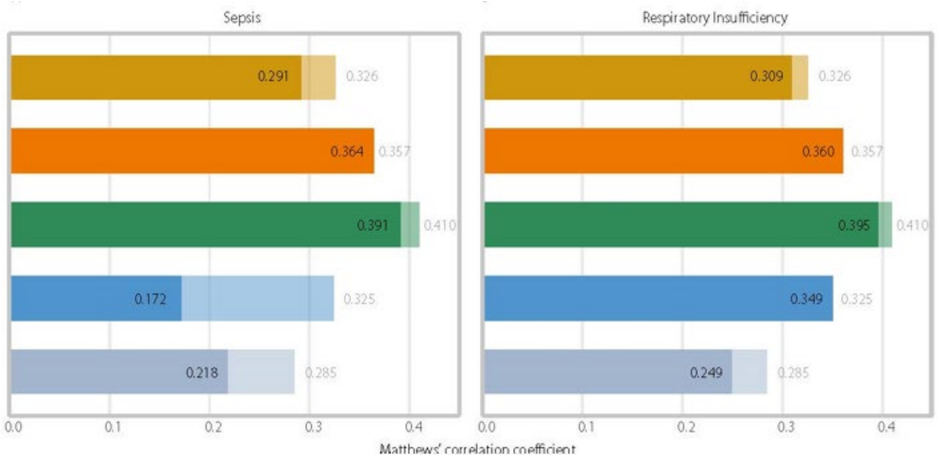




Nielsen et al., Lancet Dig Health, 2019

ICU mortality prediction performance



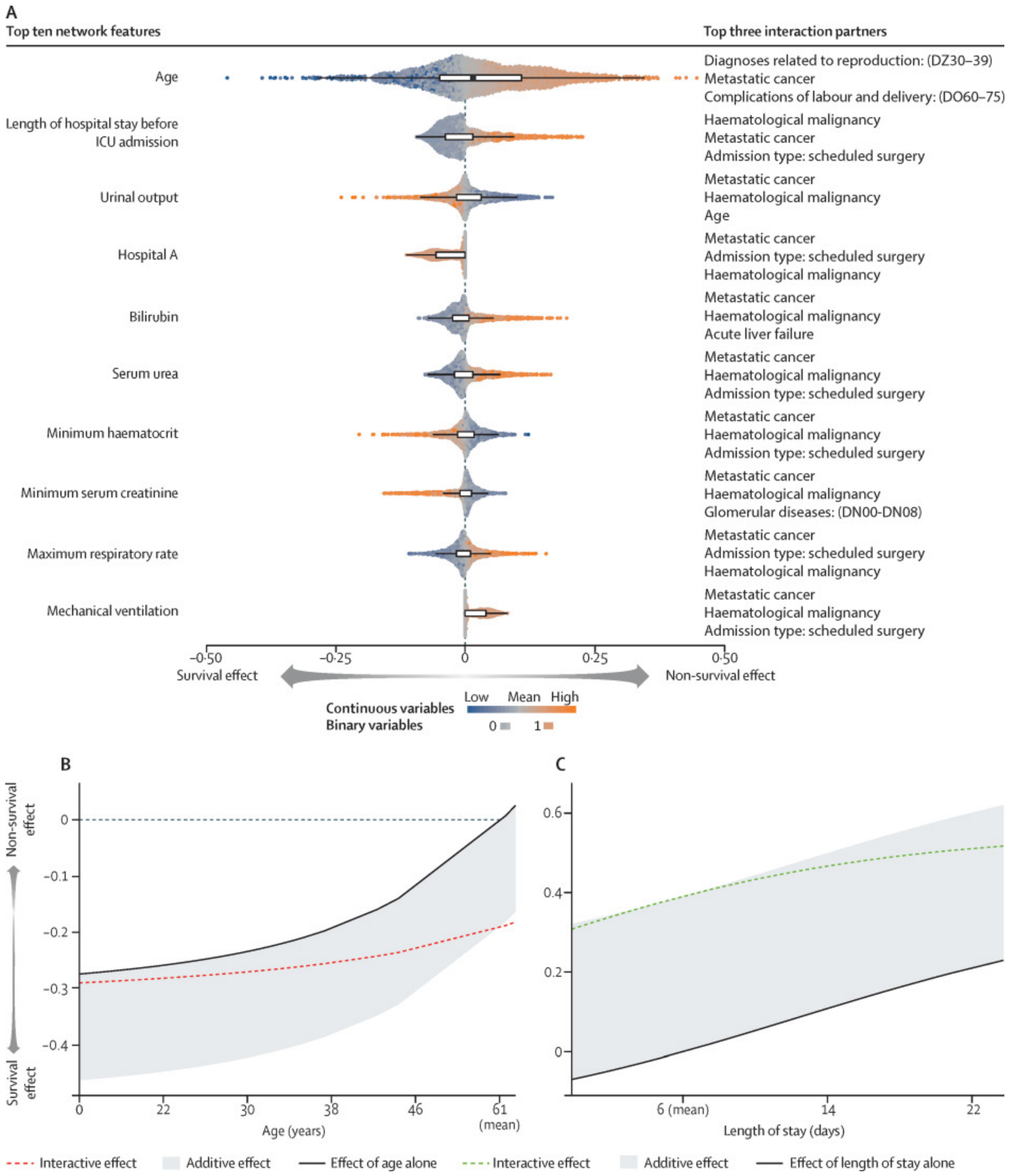


History before admission	History at admission	Aggregated history	SAPSII	APACHEII
Ten-year disease history Sex Age	Ten-year disease history Sex Age Length of stay Transfer category Hospital code	Ten-year disease history Sex Age Length of stay Transfer category Hospital code Acute physiology measures	Acute physiology measures	

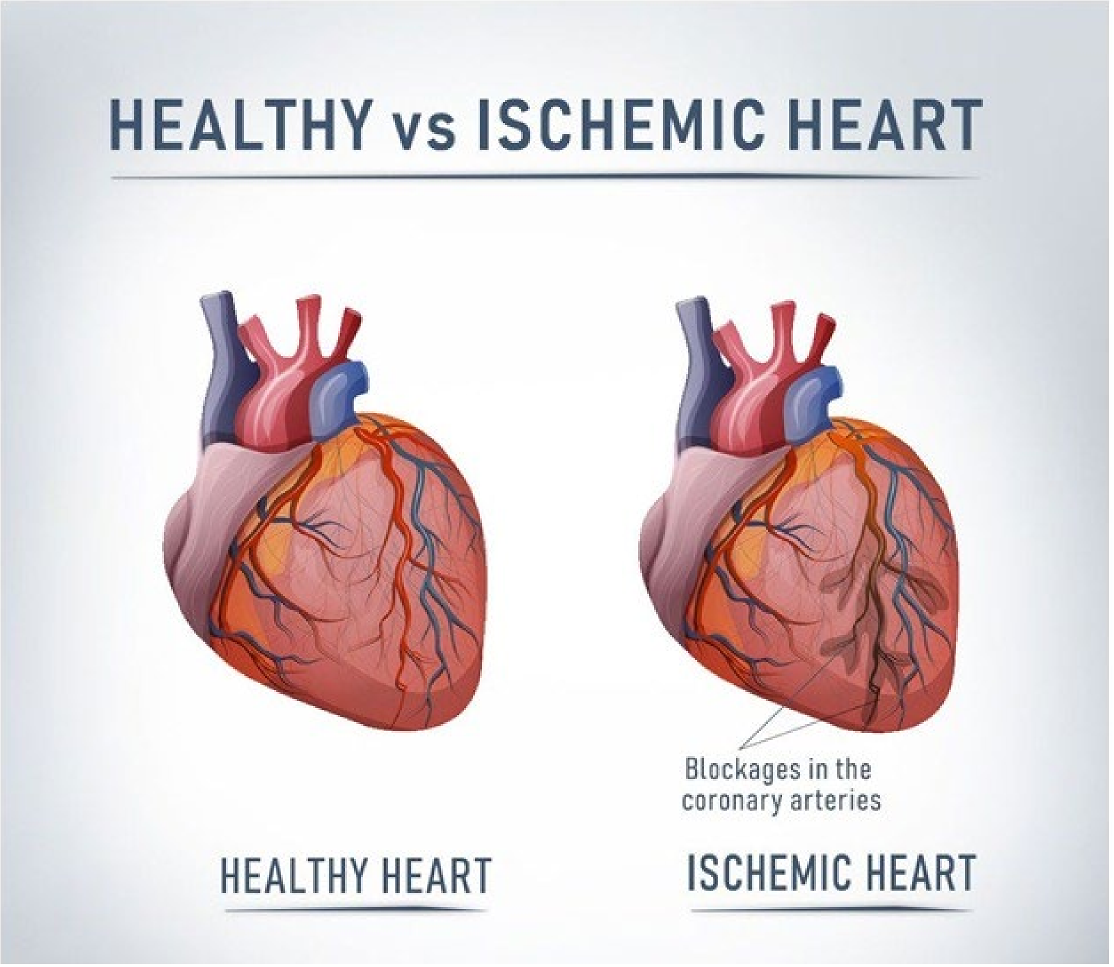
ICU mortality feature importance

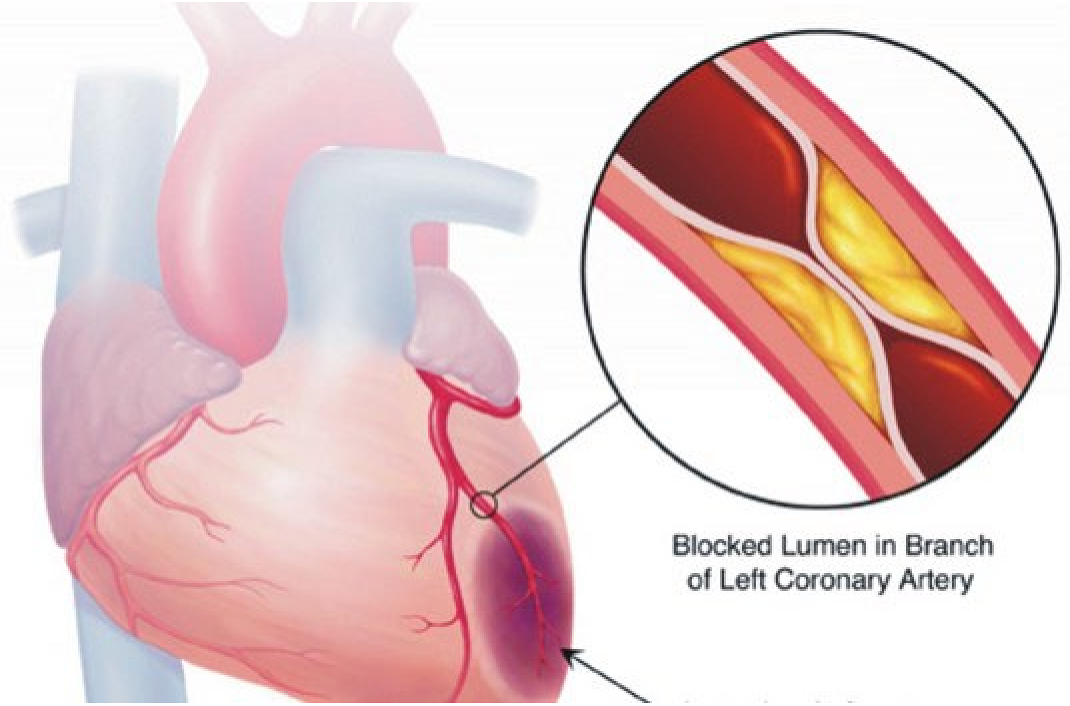
(A) Each dot one patient

Interaction between age and history of diagnoses related to reproduction (B), and interaction between length of stay before ICU admission and history of haematological malignancy (C)



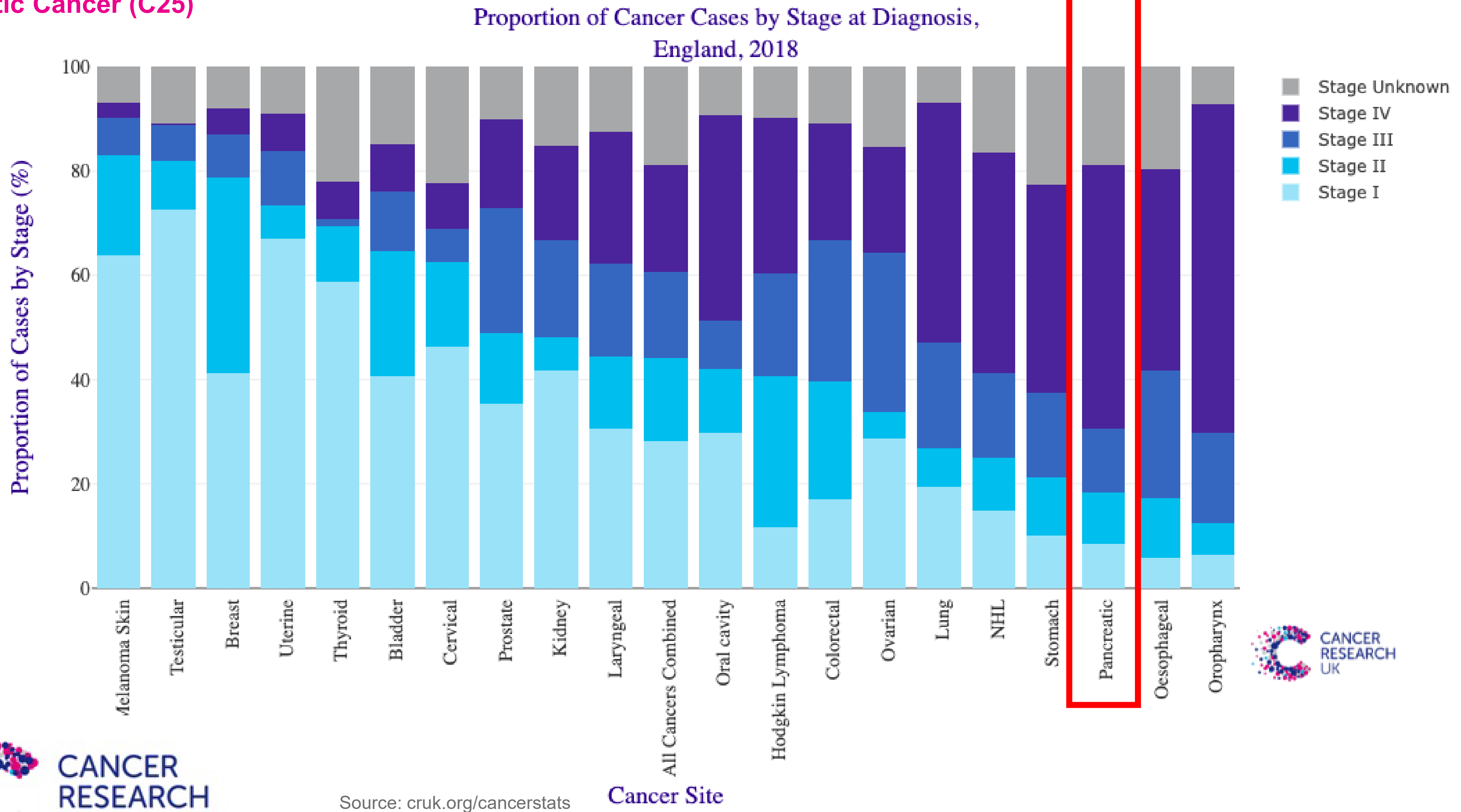
Risk prediction in ischemic heart disease

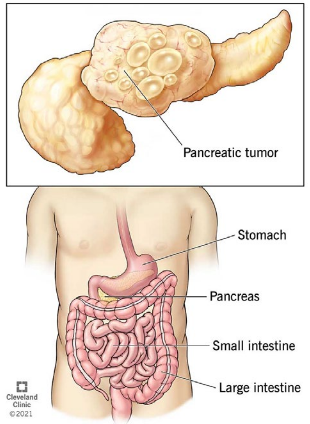




Why Pancreatic Cancer?

Proportion of Cases Diagnosed at Each Stage, All Ages
Pancreatic Cancer (C25)





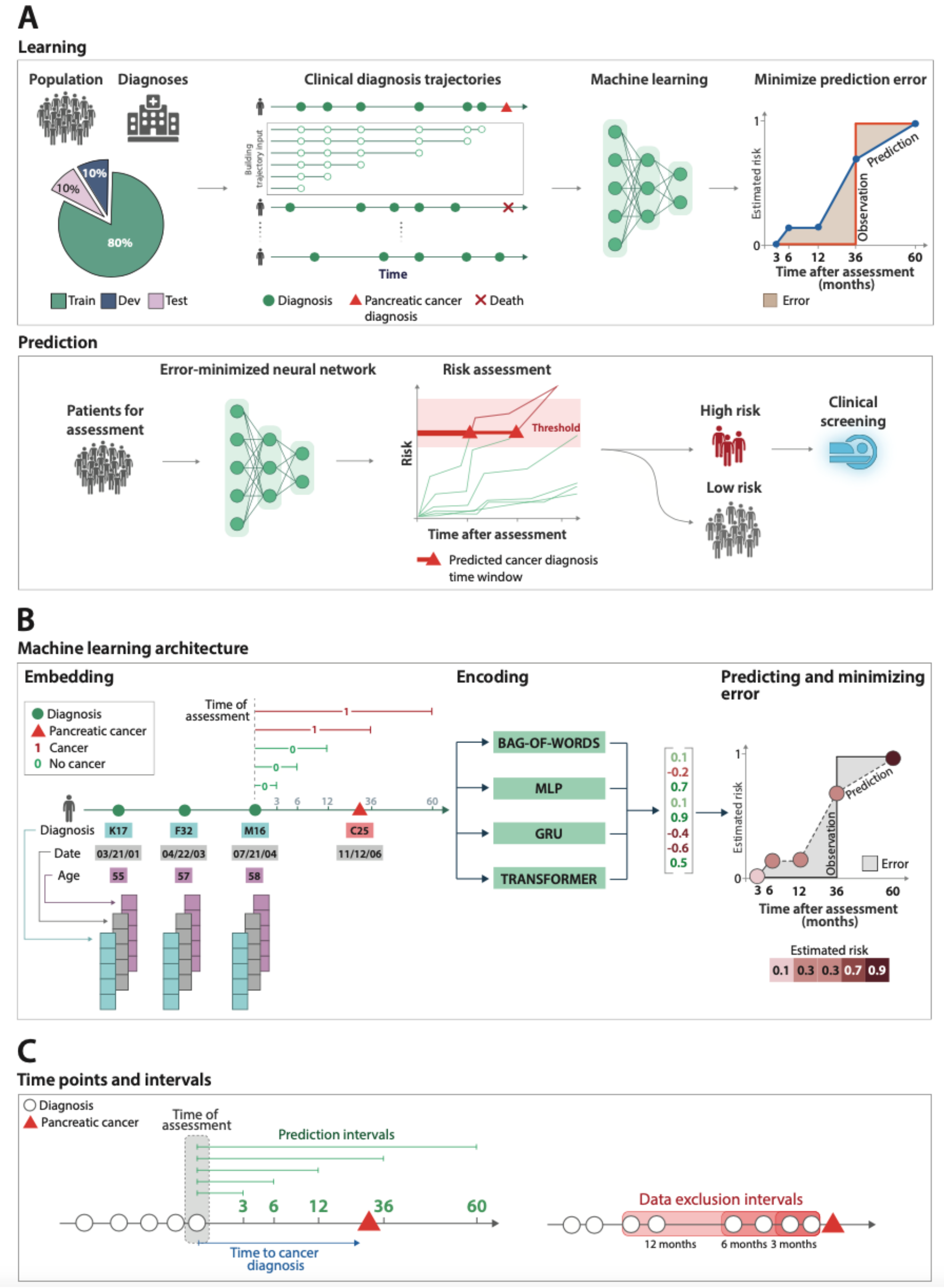
Source: cruk.org/cancerstats
You are welcome to reuse this Cancer Research UK statistics content for your own work.
Cancer Research UK

Prediction of pancreas cancer risk – training on Danish data, replication in US data

Disease histories from

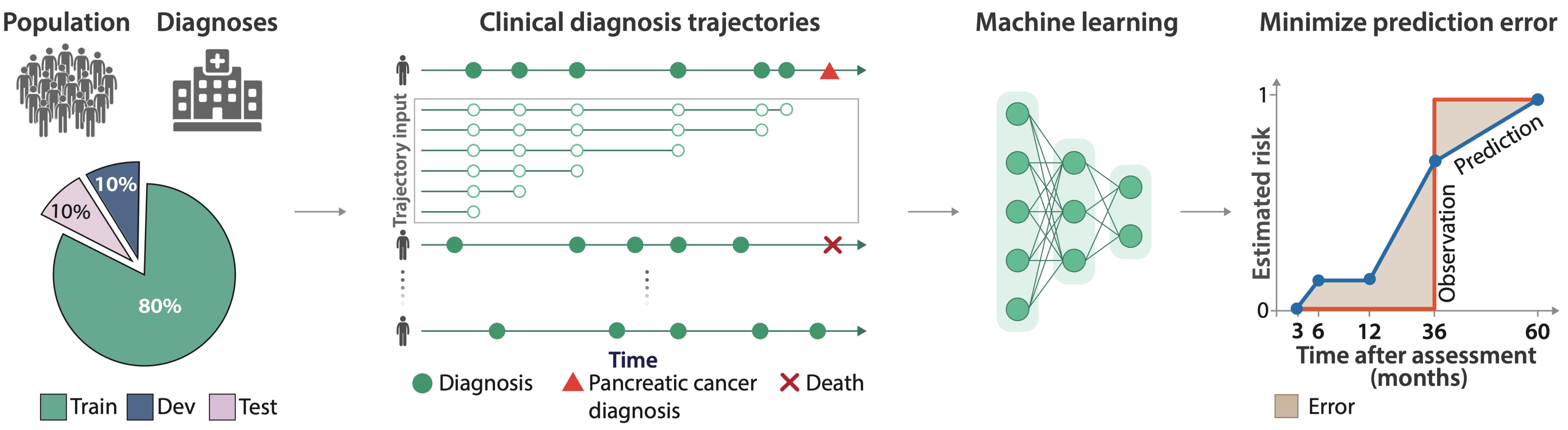
- Danish National Patient Registry (DNPR), covering 8.6 M patients between 1977-2018 (6.1 M controls, 24,000 cases, av. 23 yrs of history)

- Veteran Affairs CDW database, covering 2.9 M patients 1999-2020 (1.9 M controls, 3,800 cases, av. 12 yrs of history)

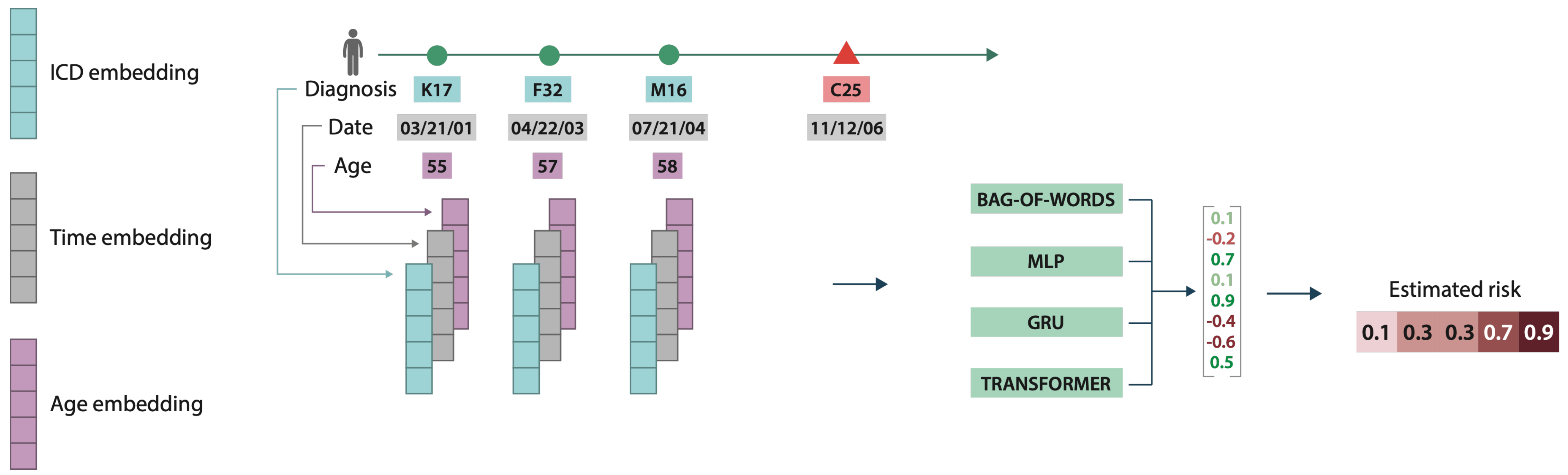


Pancreatic cancer risk predicted from disease trajectories using deep learning
Placido, Yuan, Hjaltelin, …, Brunak & Sander, Nature Medicine 2023

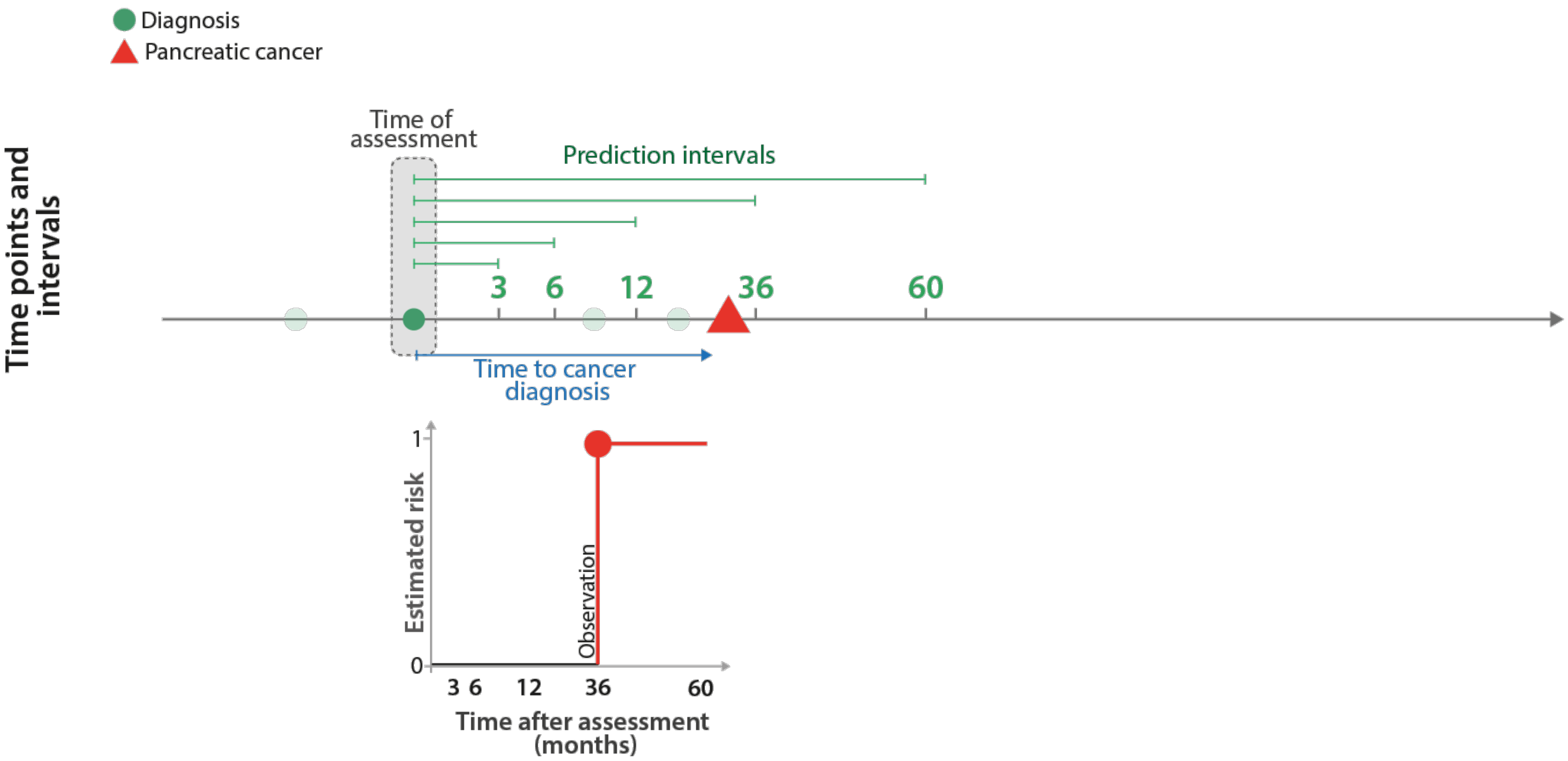
Training, development validation, test



Input data encoding: – diagnosis trajectory, dates and age

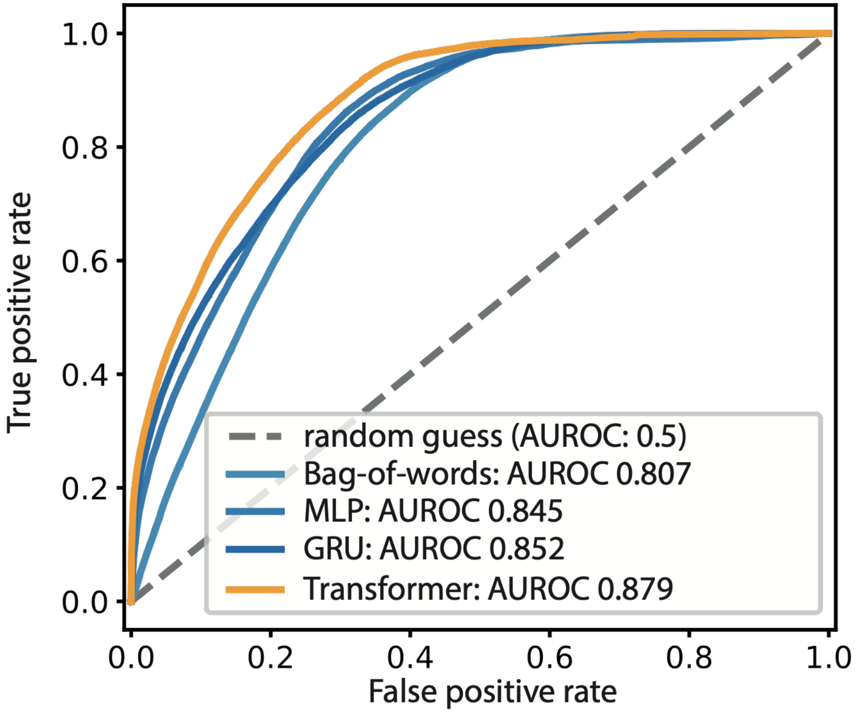


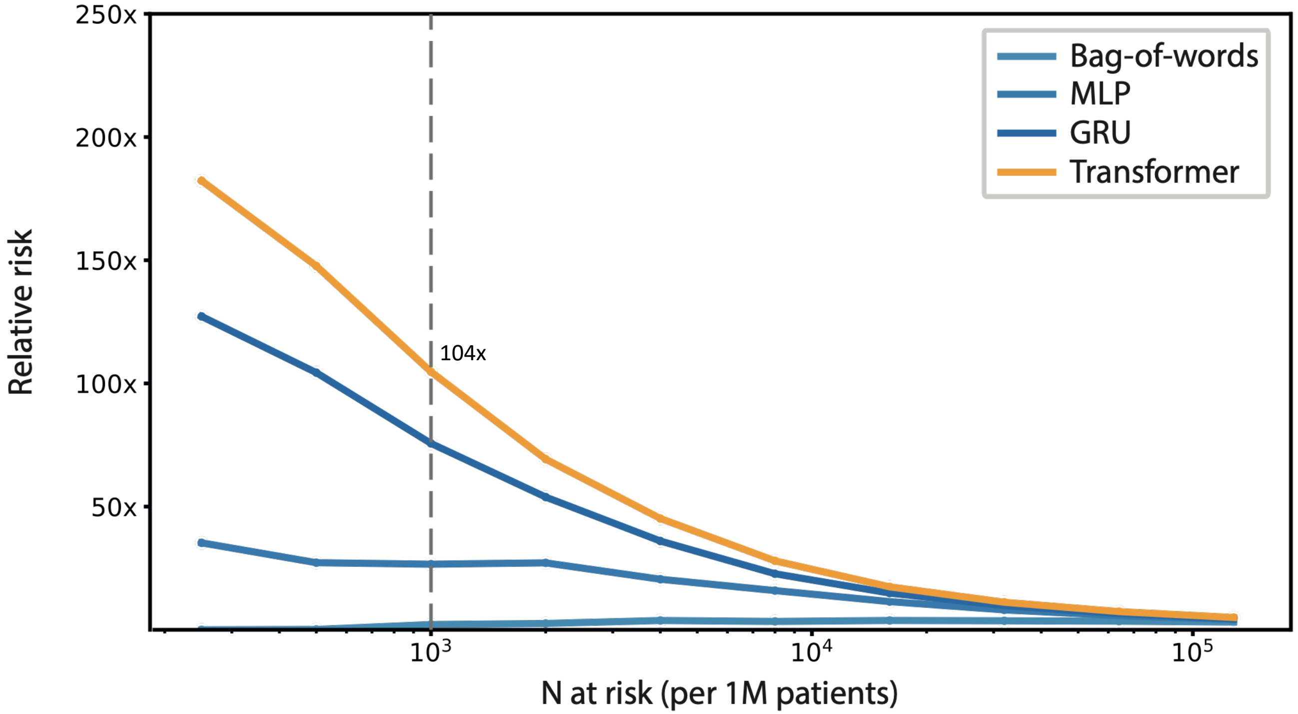
Output modelling



Results - Model comparison

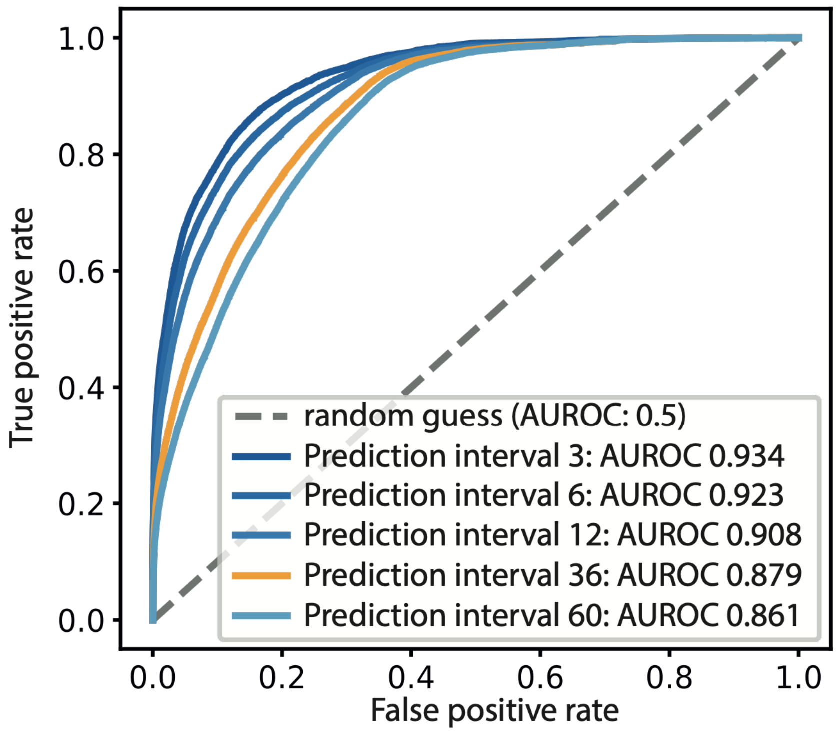
Clinical utility assessed with AUROC and Relative Risk

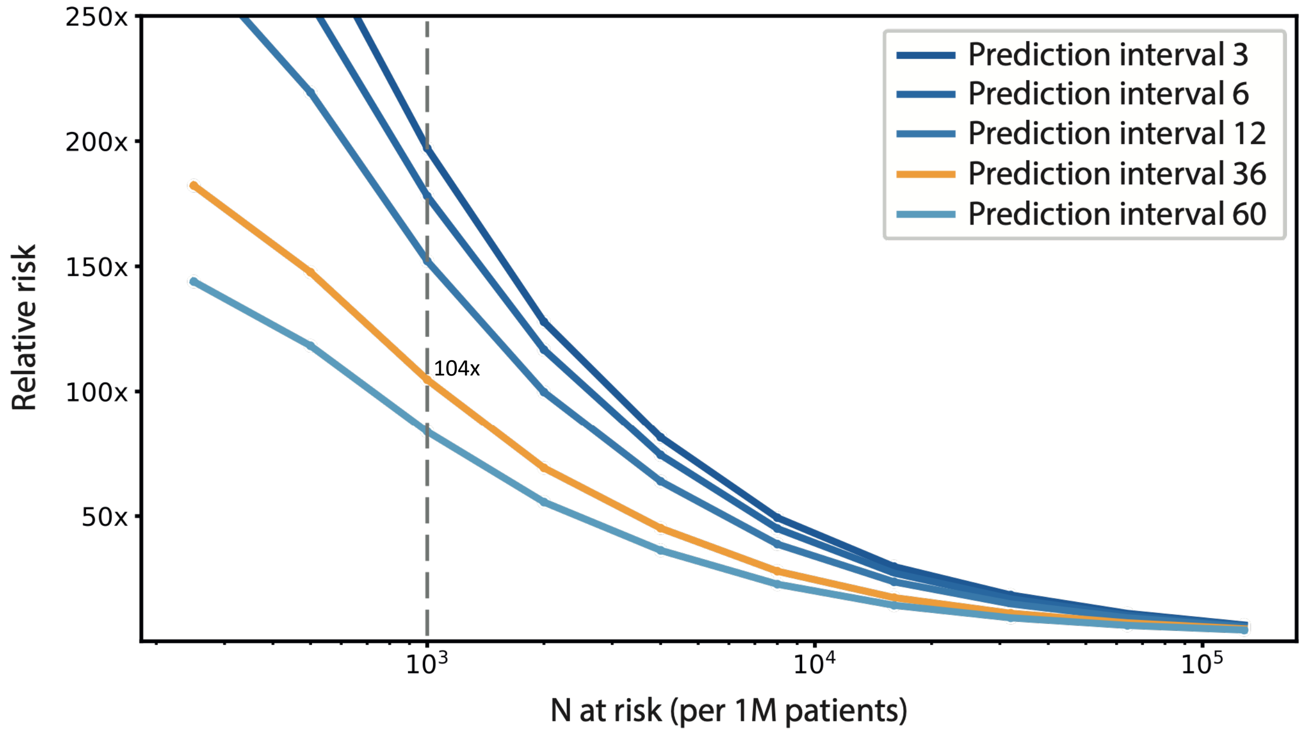




Results - Prediction intervals

Clinical utility assessed with AUROC and Relative Risk



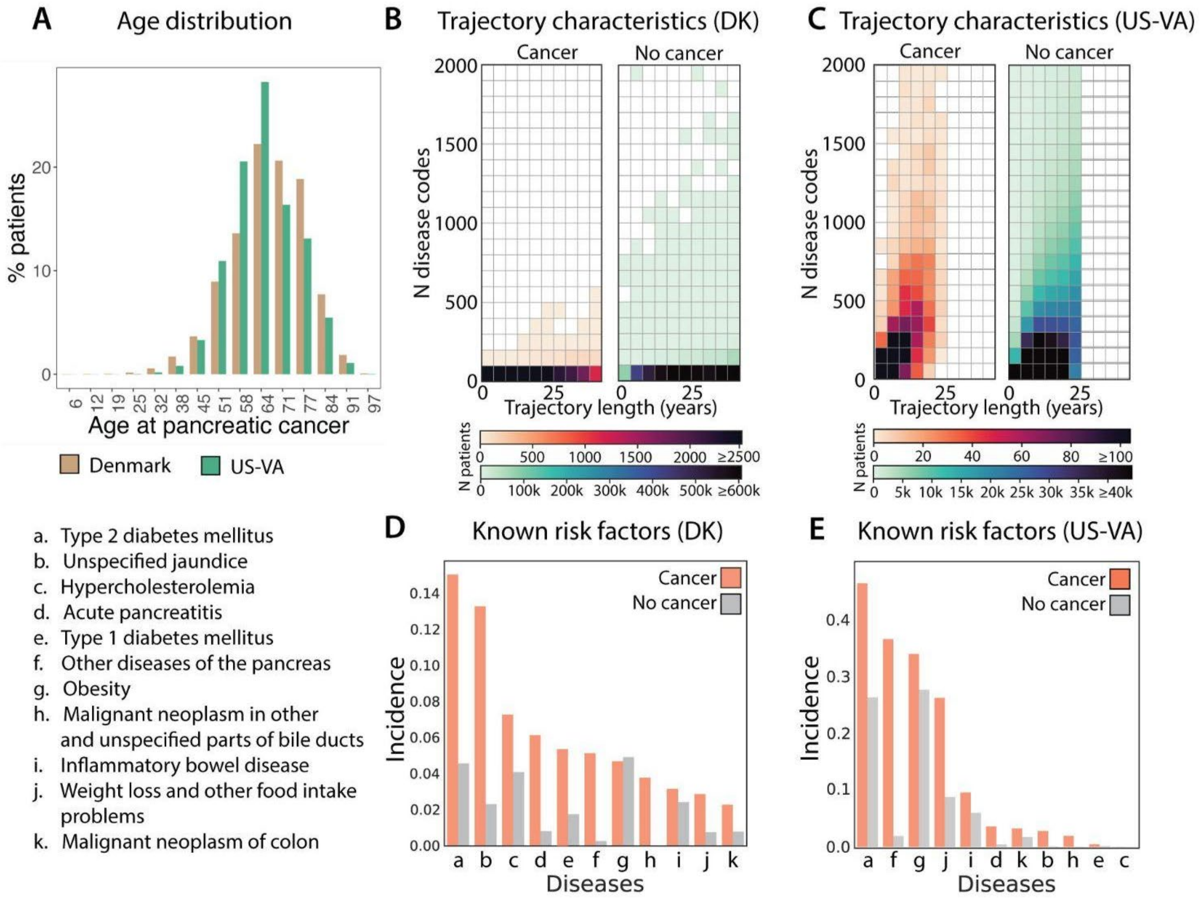


Different Denmark & US VA EHR features

Characteristics of Danish and US-VA dataset

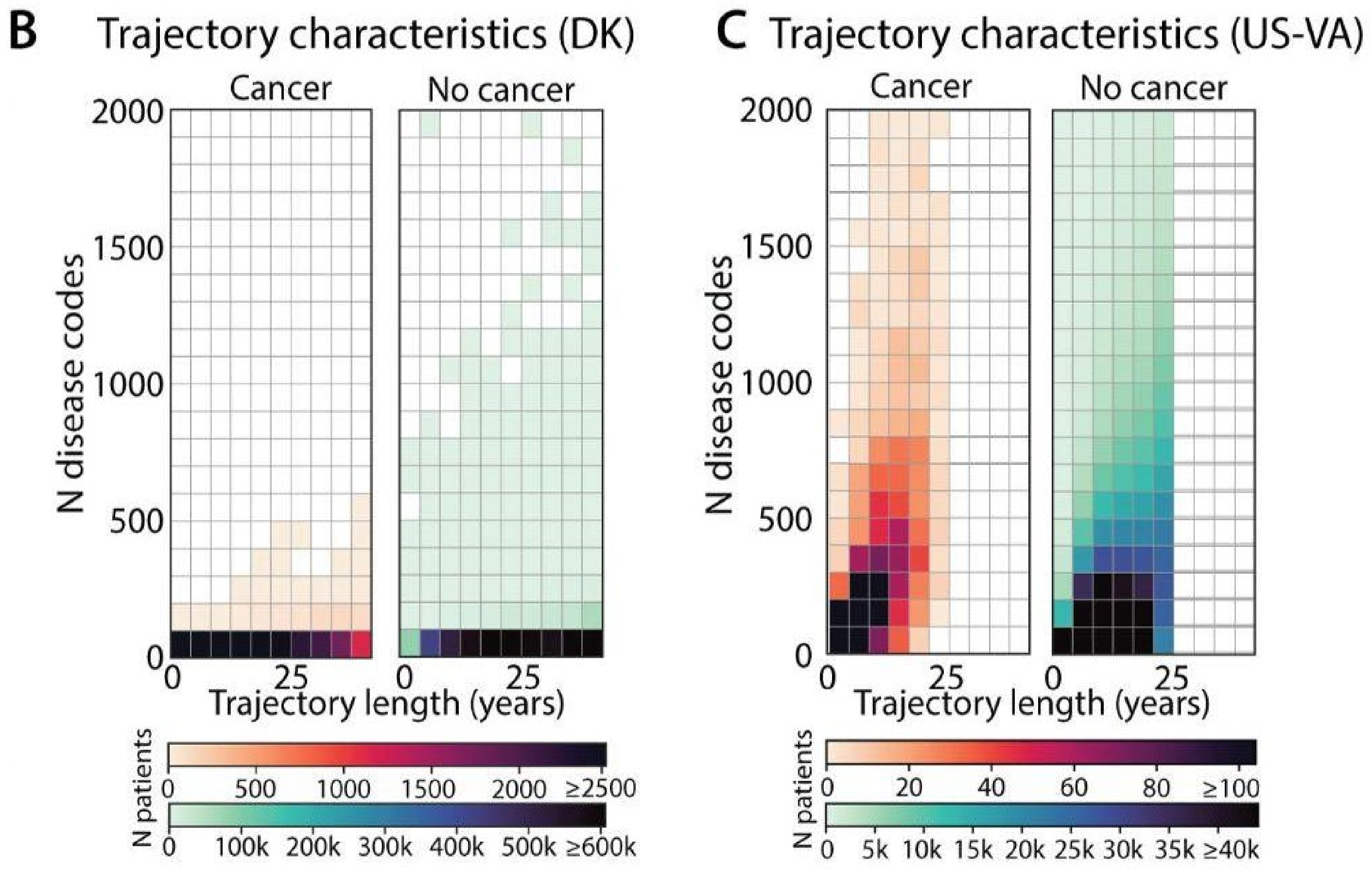
General cohort information	Danish dataset	US-VA dataset
Dataset timeline	1977-2018	1999-2020
Total N patients	8,110,706	2,962,383
Male (%)	4,030,504 (49.7%)	2,538,762 (85.7%)
Female (%)	4,080,202 (50.3%)	423,621 (14.3%)
Median N disease codes per patient	22	188
Median length of trajectory in years	23.0	12.0
PC cohort information		
Total N patients	23,985	3,869
Male (%)	11,880 (49.5%)	3,741 (96.7%)
Female (%)	12,105 (50.5%)	128 (3.3%)
Median N disease codes per patient	18	121
Median length of trajectory in years	17.0	8.0
Median age at PC diagnosis	70.0	68.0
N disease codes 3 months pre-PC	95,358	368,295
N disease codes 6 months pre-PC	27,131	535,631
N disease codes 12 months pre-PC	38,109	818,522
N disease codes >12 months pre-PC	480,830	3,469,239

Abbreviations: PC: pancreatic cancer.



a. Type 2 diabetes mellitus
b. Unspecified jaundice
c. Hypercholesterolemia
d. Acute pancreatitis
e. Type 1 diabetes mellitus
f. Other diseases of the pancreas
g. Obesity
h. Malignant neoplasm in other and unspecified parts of bile ducts
i. Inflammatory bowel disease
j. Weight loss and other food intake problems
k. Malignant neoplasm of colon

Different Denmark & US VA EHR coding features



Many more codes in the US data per patient than in the DK data

Pancreatic cancer risk predicted from disease trajectories using deep learning
Placido, Yuan, Hjaltelin, …, Brunak & Sander, Nature Medicine 2023

Feature importance ranking using explainability methods

C

Feature contributions - No exclusion (DK)

	Cancer in 0-6 months	Cancer in 6-12 months	Cancer in 12-24 months	Cancer in 24-36 months
1	Unspecified jaundice	Other diseases of biliary tract	Medical observation and evaluation for suspected diseases and conditions	Medical observation and evaluation for suspected diseases and conditions
2	Medical observation and evaluation for suspected diseases and conditions	Unspecified jaundice	Other diseases of biliary tract	Other diseases of pancreas
3	Other diseases of biliary tract	Medical observation and evaluation for suspected diseases and conditions	Other diseases of pancreas	Other diseases of biliary tract
4	Abdominal and pelvic pain	Other diseases of pancreas	Abdominal and pelvic pain	Non-insulin-dependent diabetes mell itus
5	Malignant neoplasm of other and uns pecified parts of biliary tract	Malignant neoplasm of other and uns pecified parts of biliary tract	Non-insulin-dependent diabetes mell itus	Unspecified jaundice
6	Other diseases of pancreas	Abdominal and pelvic pain	Malignant neoplasm of other and uns pecified parts of biliary tract	Abdominal and pelvic pain
7	Secondary malignant neoplasm of res piratory and digestive organs	Secondary malignant neoplasm of res piratory and digestive organs	Unspecified jaundice	Malignant neoplasm of other and uns pecified parts of biliary tract
8	Symptoms and signs concerning food and fluid intake	Non-insulin-dependent diabetes mell itus	Other functional intestinal disorde rs	Gastritis and duodenitis
9	Non-insulin-dependent diabetes mell itus	Malignant neoplasm without specific ation of site	Diseases of pancreas	Insulin-dependent diabetes mellitus
10	Other anaemias	Other anaemias	Secondary malignant neoplasm of res piratory and digestive organs	Other anaemias

D

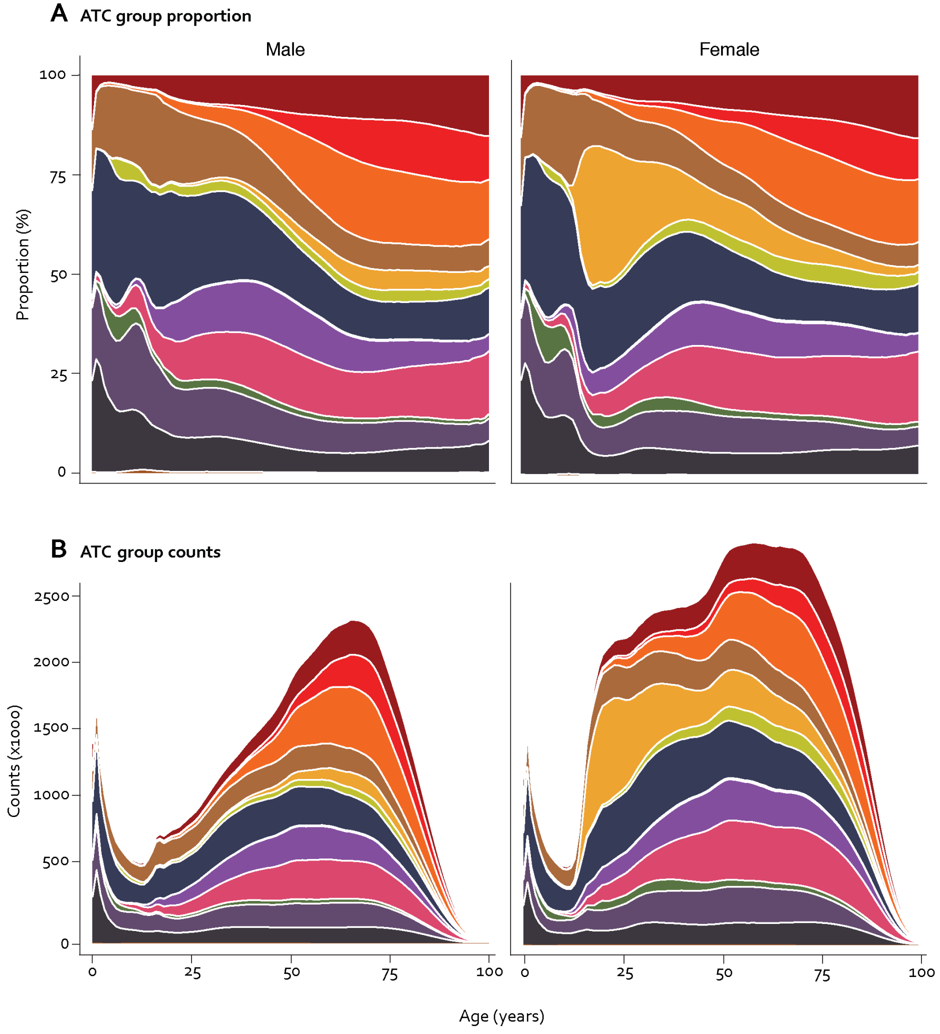
Feature contributions - No exclusion (US-VA)

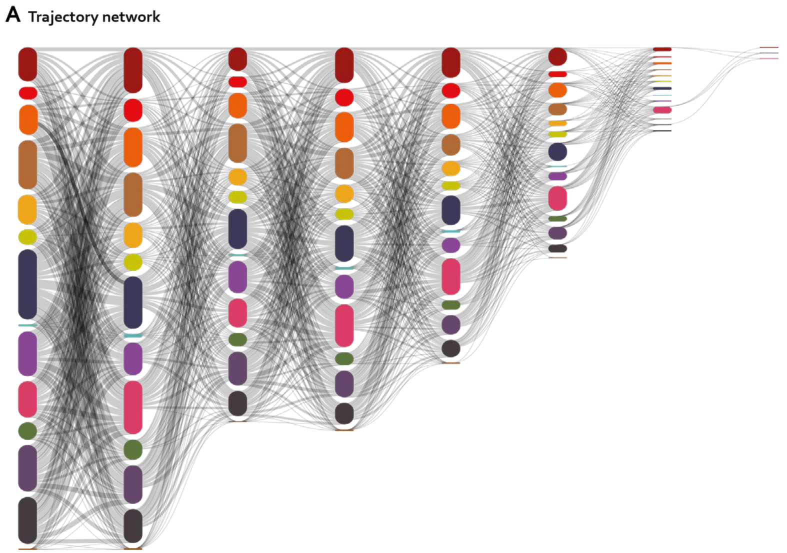
	Cancer in 0-6 months	Cancer in 6-12 months	Cancer in 12-24 months	Cancer in 24-36 months
1	Acute pancreatitis	Acute pancreatitis	Abdominal and pelvic pain	Diabetes mellitus
2	Abdominal and pelvic pain	Diabetes mellitus	Other diseases of biliary tract	Other diseases of liver
3	Other diseases of biliary tract	Other diseases of biliary tract	Diabetes mellitus	Persons encountering health services in other circumstances
4	Diabetes mellitus	Symptoms and signs concerning food and fluid intake	Persons encountering health services in other circumstances	Abdominal and pelvic pain
5	Other diseases of pancreas	Persons encountering health services in other circumstances	Acute pancreatitis	Other diseases of biliary tract
6	Symptoms and signs concerning food and fluid intake	Malignant neoplasm of trachea, bronchus or lung	Dependence of opioids, sedatives, cocaine, cannabinoids, hallucinogens, or other psychoactive substances	Nausea and vomiting
7	Disorders of social functioning with onset specific to childhood and adolescence	Abdominal and pelvic pain	Abuse of alcohol, tobacco, opioids, sedatives, cocaine, cannabinoids, hallucinogens, or other psychoactive substances	Abuse of alcohol, tobacco, opioids, sedatives, cocaine, cannabinoids, hallucinogens, or other psychoactive substances
8	Essential (primary) hypertension	Other diseases of pancreas	Cough, haemorrhage from respiratory passages	Unspecified jaundice, or skin eruption
9	Persons encountering health services in other circumstances	Dependence of opioids, sedatives, cocaine, cannabinoids, hallucinogens, or other psychoactive substances	Secondary malignant neoplasm of respiratory and digestive organs	Cataract
10	Examination and observation for other reasons	Other dermatitis	Cataract	Dependence of opioids, sedatives, cocaine, cannabinoids, hallucinogens, or other psychoactive substances

ICD-10 chapters

- II Neoplasms
- III Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
- IV Endocrine, nutritional and metabolic diseases
- V Mental and behavioral disorders
- VII Diseases of the eye and adnexa
- IX Diseases of the circulatory system
- XI Diseases of the digestive system
- XII Diseases of the skin and subcutaneous tissue
- XVIII Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
- XXI Factors influencing health status and contact with health services

ATC drug groups in 1.1 billion male and female GP prescriptions according to age





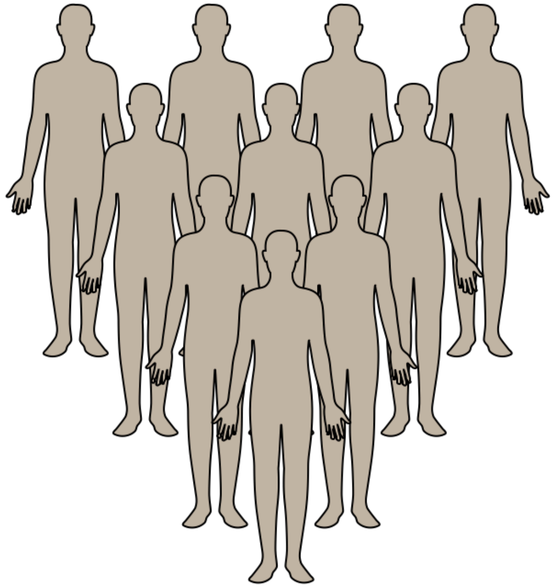
ATC group

- A alimentary tract and metabolism
- B blood and blood forming organs
- C cardiovascular system
- D dermatologicals
- G genito-urinary system and sex hormones
- H systemic hormonal preparations
- J antiinfectives for systemic use
- L antineoplastic and immunomodulating agents
- M musculoskeletal system
- N nervous system
- P antiparasitic products, insecticides and repellents
- R respiratory system
- S sensory organs
- V various

Aguayo Orozco et al.,
Npj Digital Medicine, 2021

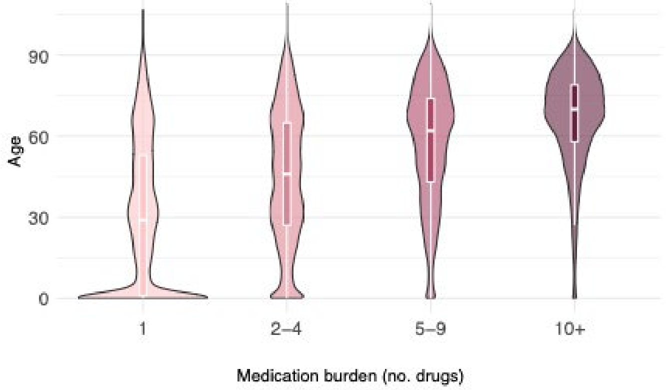
Dosage-trajectories in in-hospital polypharmacy analysis

All inpatient admissions in Capital Region of Denmark (12 hospitals), 2008-2016



N patients	1,069,873
N admissions	3,161,647 (54% F)
N drug prescriptions	24,379,285
N drugs, median (IQR)	6 (3-11)
Age, median (IQR)	59 (36-73)

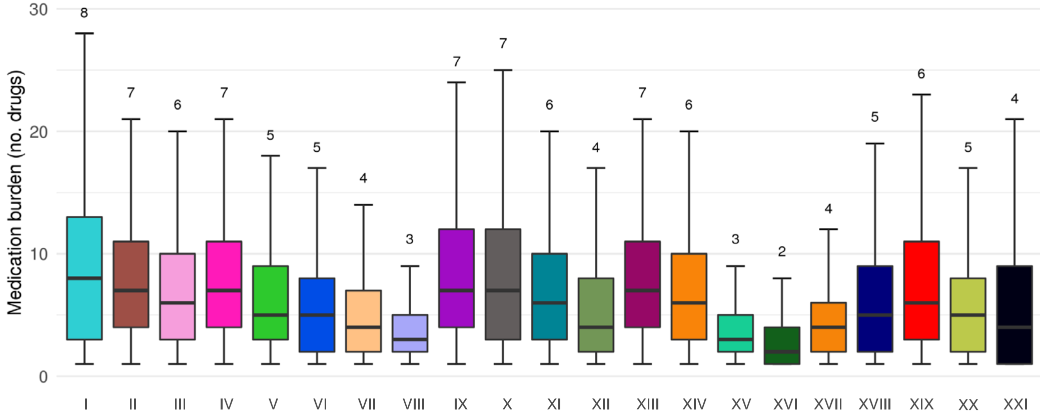
Polypharmacy is positively correlated with age (Pearson ρ:0.40)



The degree of polypharmacy varies across different primary diagnoses

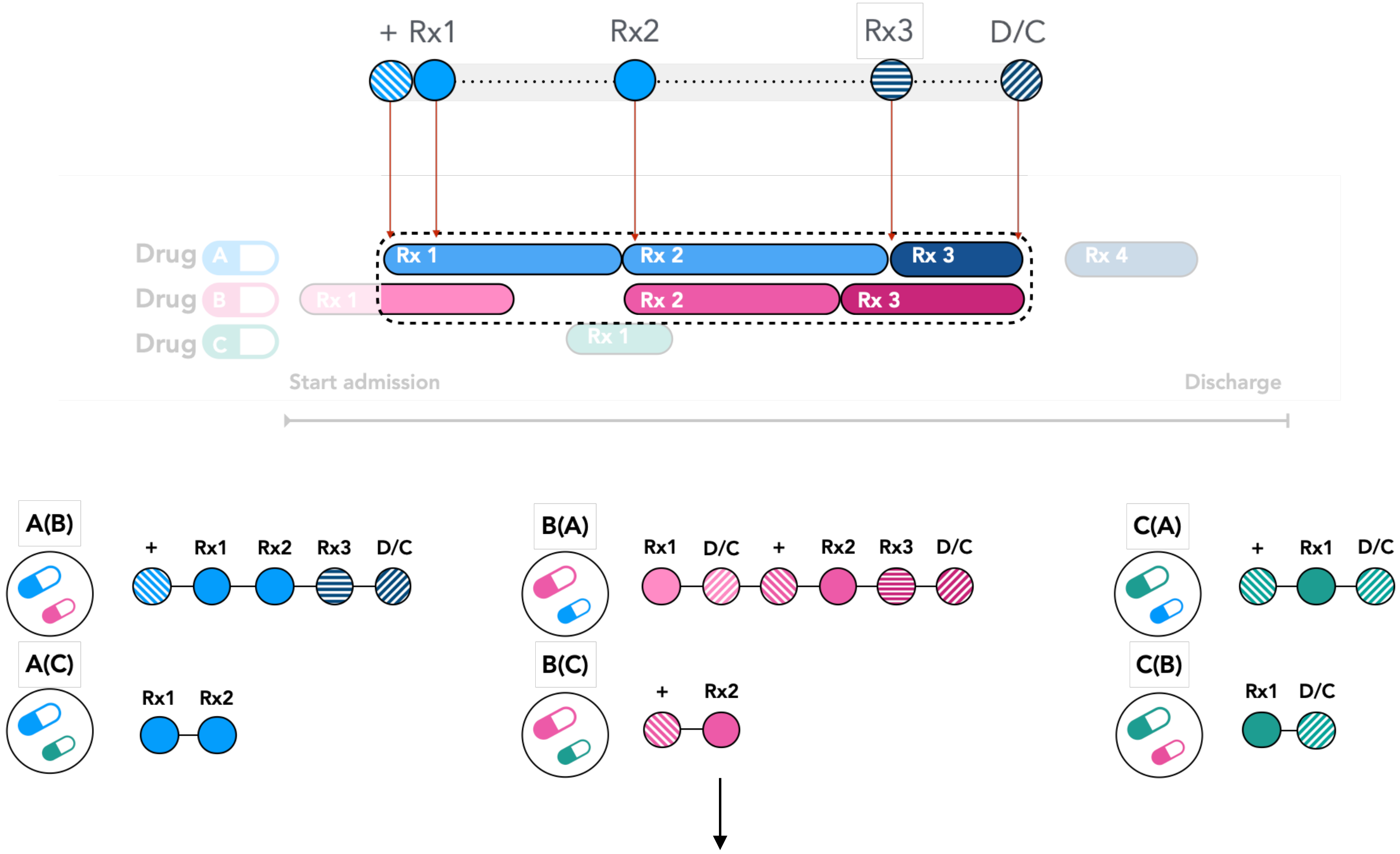
+ I: Infections
IX: Cardiovascular
X: Respiratory diseases

- XV, XVI: Pregnancy and perinatal
XVII: Congenital diseases
VIII: Ear and mastoid processes



Leal et al.
Npj Digital Medicine, 2023

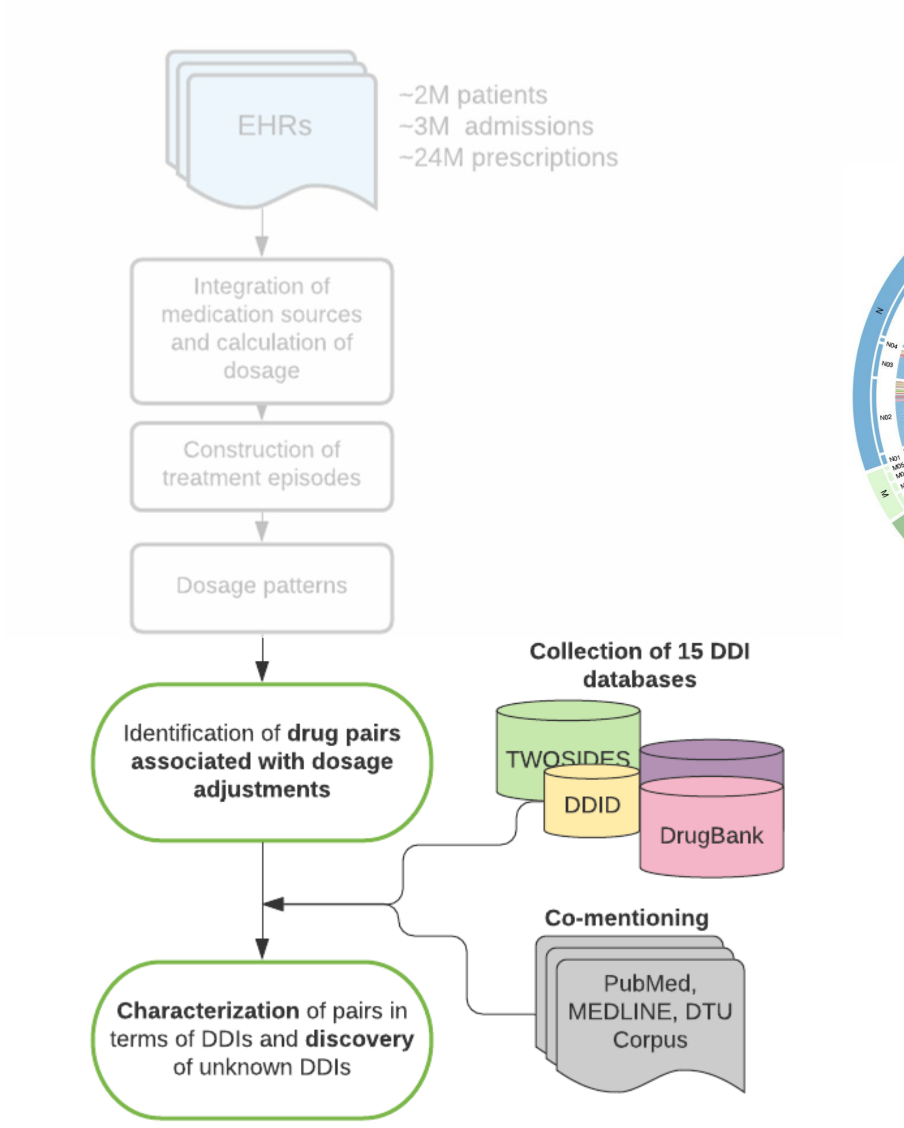
From 185 M treatment episodes to co-medication pairs

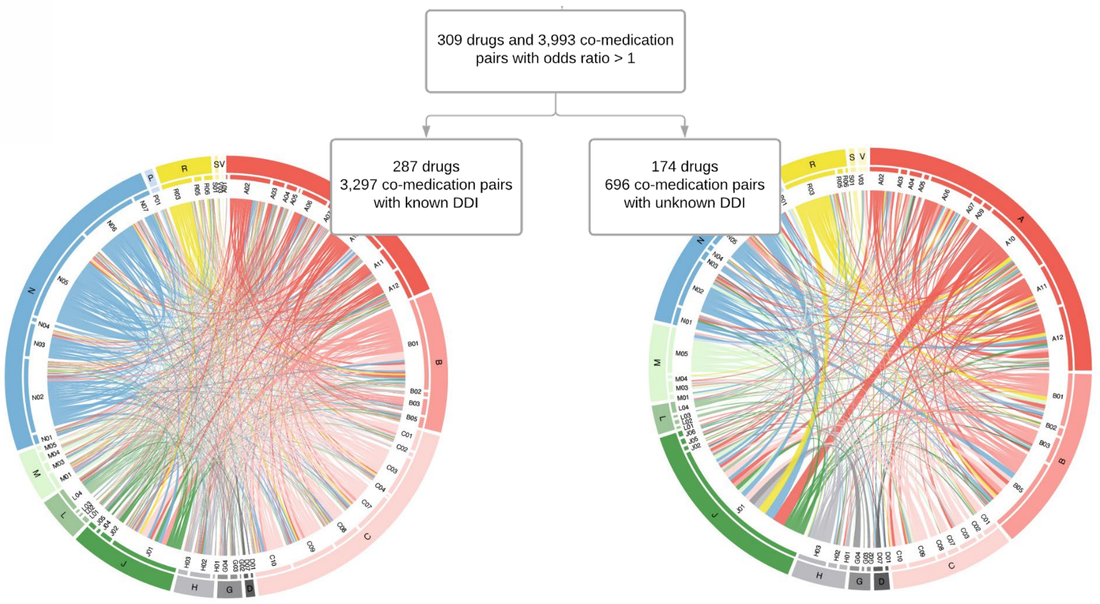


185M concomitant treatment episodes

413 index drugs: 77,494 co-medication pairs ⟶ 3,993 pairs with significant dosage changes

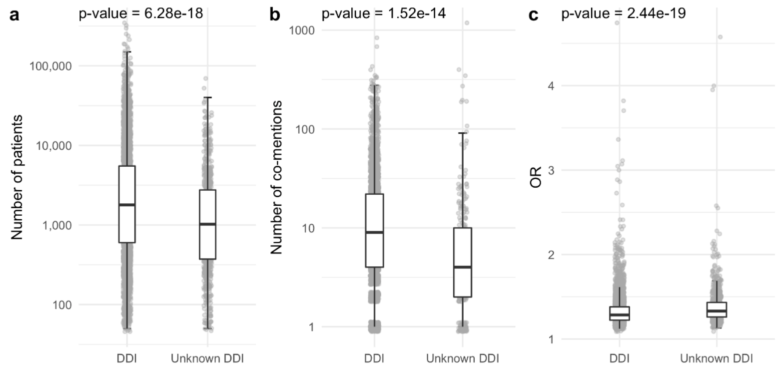
83% of the co-medication pairs with significant dosage changes are associated to known Drug-Drug Interactions





Increased proportion of antibiotics and musculoskeletal drugs, lower in nervous and cardiovascular system drugs

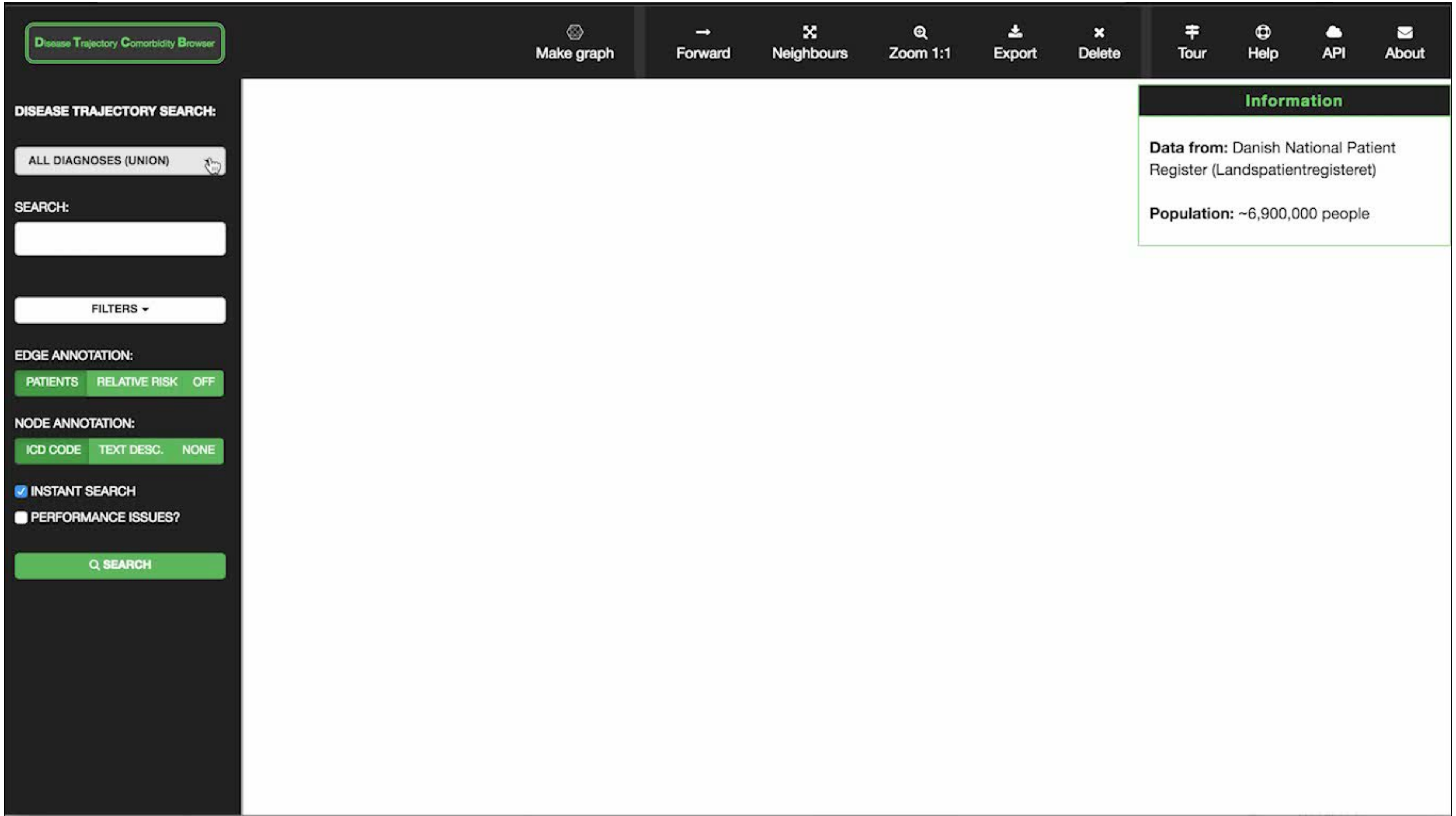
696 pairs within known DDIs had lower patient volume, were less described together in the literature and had higher ORs



Leal et al.
Npj Digital Medicine, 2023

The Danish Disease Trajectory Browser:
http://dtb.cpr.ku.dk

Siggaard et al., Nature Comm, 2020



Δ population-wide health and deep learning models

- Health data driven:
  - Redefine phenotypes as trajectories
  - Re-assign patients to the proper sub-category
  - Enable prediction using trajectories
  - Handle life long data capture
  - ”Live data” versus data dumps versus traditional registers
  - Progression biomarkers versus disease risk biomarkers

- Include what is not in the hospital patient records in new ways:
  - Diet
  - Genetics
  - GP events
  - Income, …
  - Education, grades in exams, …
  - Wearable data (partly EHR included)
  - Patient generated data
  - Smart meter data


Real-time registries: Danish National Patient Registry

https://quantifyresearch.com/wp-content/uploads/2022/10/LPR3-Introducing-the-new-and-improved-Danish-patient-register.pdf

Table 1 Key differences between LPR2 and LPR2

	LPR2	LPR3
Connection of health and disease course	No direct description of health and disease courses exists, instead every contact appears separately. The contact consists of patient identification, diagnoses, procedures, and related information	Additional information level which brings all contacts, diagnoses, and procedures as well as related course markers* and triggered result reports** in a coherent health and disease course, see Figure 1. For each course there is a specific course label which typifies the current independent course e.g., "COPD", "type-2 diabetes", "cancer".
Outpatient visits	All outpatient care contacts (visits) are reported as one main diagnosis code with a start and end date, and each visit in the course is only indicated with a visit date without indication of the actual duration of the visit. This means that LPR2 do not contain information on what the diagnosis is on a given visit date.	Every contact between the patient and the health care system is independent such as diagnostic, observation, treatment, counselling and is registered independently with the exact time of the visit (down to seconds) including visit-specific diagnoses and/or procedure codes (but still with the possibility to link these elements to a specific disease course).
Code lists (different collections of SKS-codes)	Diagnoses, operations, and treatments are reported based on the health care classification system (SKS).	In LPR3, the framework for which SKS codes can be used in reporting is based on code lists. This list is continuously updated with new codes when needed. The code lists place restriction on the use of SKS codes e.g., dates of validity. The list is particularly important for the validation of the reporting, which will be more flexible and easier to maintain.
No differentiation between in-patient and outpatient visit	Type of patient contact is reported as outpatient, emergency or in-patient.	This is not distinguished in LPR3. Instead, the duration of each contact is registered along with information about type of contact (physical, virtual, external contact, death, or diagnosis recording). By use of this information, it is still possible to make the distinction comparable to the LPR2 patient type if needed.
Public and private hospitals	The content of the data is not the same for the public and private hospitals. It depends on whether it is reported directly to LPR2 or via the so called "MiniPas".	The reporting requirements are now identical for the public and the private clinics including data from psychiatry.
Additional mandatory codes	Mandatory to add SKS codes in connection with reporting of supplementary information (e.g., for births and cancer) and various types of markers (e.g., for waiting time and package offers).	The mandatory additional information is generally built in as fixed elements in the reporting (e.g., as results in registration of results or as course markers)**. There is no requirement for additional coding in LPR3, but it is possible to report additional codes for diagnoses and procedures.

* Real time markers based on legislation and political decisions about registration e.g., packages offer in Cancer treatment
** Results reports are triggered by course markers, contacts, diagnoses, and procedures and forwards the mandatory reports and notifications.

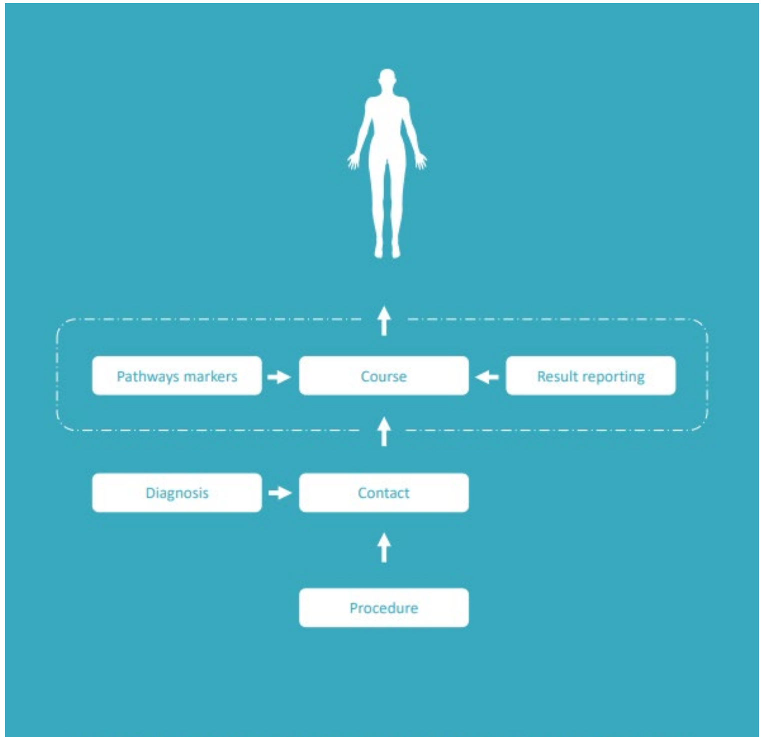


Figure 1 Concept model of LPR3 including the additional level. Recreated from "LPR-indberetningsve

Real Time

Acknowledgements

Trajectory registry data analysis

Anders Boeck Jensen, CPR/KU, Christian Simon, CPR/KU
Karina Basanik, CPR/KU, Mette Beck, CPR/KU
Robert Eriksson, CPR/KU**, Isabella Friis Jørgensen, CPR**
David Westergaard, CPR, Jessica Hu Hjaltelin, CPR
Troels Siggaard, CPR, Alejandro Aguayo Orozco, CPR
Amalie Dahl Haue, CPR Pope Mosely, CPR/U. New Mexico
Lars Juhl Jensen, CPR

Pancreas cancer

Davide Placido, Bo Yuan, Jessica X. Hjaltelin, Chunlei Zheng,
Amalie D. Haue, Piotr J. Chmura, Chen Yuan, Jihye Kim,
Renato Umeton, Gregory Antell, Alexander Chowdhury,
Alexandra Franz, Lauren Brais, Elizabeth Andrews, Debora S. Marks,
Aviv Regev, Siamack Ayandeh, Mary T. Brophy, Nhan V. Do,
Peter Kraft, Brian M. Wolpin, Michael H. Rosenthal,
Nathanael R. Fillmore**, Chris Sander**

Ischemic heart disease

Peter C. Holm, Amalie D. Haue, David Westergaard, Timo Röder, Karina Banasik,
Vinicius Tragante, **Alex H. Christensen,** Laurent Thomas, Therese H. Nøst, Anne-Heidi Skogholt,
Kasper K. Iversen, Frants Pedersen, Dan E. Høfsten, Ole B. Pedersen, Sisse Rye Ostrowski,
Henrik Ullum, Mette N. Svendsen, Iben M. Gjødsbøl, Mette Gørtz, Mette Hartlev,
Thorarinn Gudnason, Daníel F. Guðbjartsson, Anna Helgadottir, Kristian Hveem, Lars V. Køber,
Hilma Holm, Kari Stefansson, **Henning Bundgaard**

ICU mortality prediction

Annelaura B Nielsen, Hans-Christian Thorsen-Meyer,
Kirstine Belling, Anna P Nielsen, Cecilia E Thomas,
Piotr J Chmura, Mette Lademann, Pope L Moseley,
Marc Heimann, Lars Dybdahl, Lasse Spangsege,
Patrick Hulsen, **Anders Perner**